MATHEMATICAL MODELING OF MULTI-LEVEL BEHAVIOR OF THE EMBRYONIC STEM CELL SYSTEM DURING SELF-RENEWAL AND DIFFERENTIATION

by

Keith Daniel Task

B.S. Chemical Engineering, University of Pittsburgh, 2005

Submitted to the Graduate Faculty of

The Swanson School of Engineering in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2014

UNIVERSITY OF PITTSBURGH

SWANSON SCHOOL OF ENGINEERING

This dissertation was presented

by

Keith Daniel Task

It was defended on

July 7, 2014

and approved by

William Federspiel, PhD, Professor, Department of Bioengineering

Louis Luangkesorn, PhD, Research Assistant Professor, Department of Industrial Engineering

Robert Parker, PhD, Associate Professor, Department of Chemical Engineering

Dissertation Director: Ipsita Banerjee, PhD, Assistant Professor, Department of Chemical Engineering

MATHEMATICAL MODELING OF MULTI-LEVEL BEHAVIOR OF THE EMBRYONIC STEM CELL SYSTEM DURING SELF-RENEWAL AND DIFFERENTIATION

Keith Daniel Task, PhD

University of Pittsburgh, 2014

Embryonic stem cells (ESC) are pluripotent cells derived from the inner cell mass of the blastocyst. These cells have the unique properties of unlimited self-renewal and differentiation capability. ESC therefore hold huge potential for use in therapeutic applications in regenerative medicine. This potential has been demonstrated *in vitro* by directing differentiation of ESC to various cell types by modulating the soluble and insoluble cues to which the cells are exposed. Despite their great potential, current differentiation methods are still limited in the yield and functionality of the ESC-derived mature phenotype. We hypothesize the lack of mechanistic understanding of the complex differentiation process to be the primary reason behind their restricted success. Mathematical models, coupled to experimental data, can aid in this understanding. While the past several decades have seen advances in the mathematical analysis of biological systems, mathematical approaches to the ESC system have received limited attention. Furthermore, variability of ESC restricts direct application of deterministic approaches towards drawing mechanistic insight.

The goal of the current work is to obtain a more thorough mechanistic understanding of the ESC system through mathematical modeling. In ESC, extracellular cues guide single cell

behavior in a non-deterministic fashion, giving rise to heterogeneous populations. Therefore, in this work we focus on modeling three levels of the ESC system: intracellular, extracellular, and population. We first developed an optimization framework to identify intracellular gene regulatory interactions from time series data. We show that incorporation of the bootstrapping technique into the formulism allows for accurate prediction of robust interactions from noisy data. A regression approach was then utilized to identify extracellular substrate features influential to cellular behavior. We apply this model to identify fibrin microstructural features which guide differentiation of mESC. Finally, we developed a stochastic model to capture heterogeneous population dynamics of hESC. We demonstrate the usefulness of the model to obtain mechanistic information of cell cycle transition and lineage commitment during differentiation. Through development and utilization of different mathematical approaches to analyze multilevel behavior and variability of ESC self-renewal and differentiation, we demonstrate the applicability of mathematical models in extracting mechanistic information from the ESC system.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

First, I would like to thank my advisor, Dr. Ipsita Banerjee, for all of the guidance, advice, encouragement and knowledge which she imparted to me over the past five years. I am truly appreciative of all of the opportunities which I have had and all of the things I have learned working in your lab.

I would also like to thank the NIH (DP2-16520) and the Department of Chemical Engineering at the University of Pittsburgh for the opportunity and resources to perform this work. To everyone whom I have worked with in grad school, including my committee and collaborators, thank you for your support and ideas. I also owe a debt of gratitude to my former professors for not only a superb job at teaching me engineering, but for inspiring me to continue with my education and perform research. And to the students in the Banerjee lab: thanks for sharing an office and lab with me over the past five years. I wish you the best of luck in all that you do.

Most importantly, I would like to thank my mother, Karen, my brother Michael, and the rest of my family for their unconditional love and support. I cannot put into words what you mean to me, and without you I wouldn't have made it nearly this far in life.

I would like to dedicate this work to the memory of my father, Allen.

1.0 INTRODUCTION

1.1 EMBRYONIC STEM CELLS

Embryonic stem cells (ESC) are pluripotent cells which can give rise to any tissue type in the body. They are derived from the inner cell mass of the blastocyst, approximately 4-5 days after fertilization [1, 2]. Numerous ESC lines have been derived for numerous species, including mouse and primate [3, 4]. Human ESC (hESC) lines have also been derived, and come from unused fertilized eggs from in-vitro-fertilization (IVF) procedures [2]. If carefully maintained, these ESC can undergo unlimited self-renewal. If proper conditions are not met, these cells lose their pluripotency and become committed to a particular lineage [5]. Indeed, the purpose of stem cell research is to direct this differentiation to the desired tissue type. The first developmental stage to which the cells differentiate is that of the three primary germ layers of ectoderm, mesoderm, and endoderm. Ectoderm gives rise to tissues such as neurons and skin, mesoderm to bone and blood, and endoderm to liver and pancreas [6] (Figure 1.1). This ability of ESC to form any of these tissue types makes them a very attractive cell source for regenerative medicine.

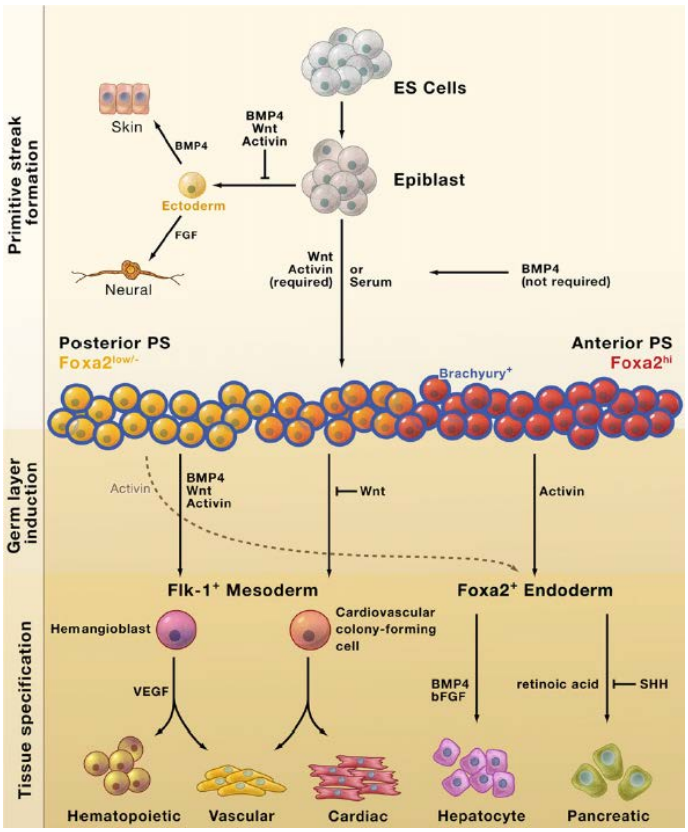

Figure 1.1. Differentiation of ESC to different cellular phenotypes

Schematic of ESC development to various lineages under different soluble cues [6]

One application that holds promise, and the focus of our current work, is the use of ESC to treat type I diabetes. This disease, in which the β-cells in the Islets of Langerhans of the pancreas are destroyed, affects more than 1.5 million Americans [7]. Because of the lack of insulin in patients with the disease, glucose uptake is impaired, which can lead to effects including hypoxia and acidosis in addition to further complications to the kidneys and cardiovascular system [8, 9]. There are several treatment options for the disease. The most common is providing the body with an exogenous insulin supply from insulin injections with glucose monitoring. The main challenge with this option is user compliance; improper injections and monitoring can lead to serious complications, including those associated with the disease

itself, such as diabetic coma, and those associated with overuse of insulin, which include hypoglycemia [9]. An alternative to exogenous insulin supply is replacement of the β-cells through islet transplantation. An advantage of this option is that if successful, patients would not need to rely on insulin injections. This procedure holds promise, as studies have reported that patients can obtain insulin independence after transplantation [10, 11]. This route also faces obstacles, the most crucial being a lack of donor organs. Because of this lack of donor cells, ESC have been looked towards as a possible source of β-cells. This involves ESC first differentiating to the endoderm germ layer, then to pancreatic and endocrine progenitors, with final maturation into insulin producing β-cells. For this to be an eventual reality in a clinical setting, ESC first have to be properly expanded in an undifferentiated state, with subsequent efficient directed differentiation to each of the developmental stages.

1.2 SELF-RENEWAL AND DIFFERENTIATION OF ESC

The two unique ESC properties of unlimited self-renewal and differentiation are equally important for therapeutic applications, and both have been demonstrated *in vitro*, mainly by modulating the external cellular microenvironment. With the trait of self-renewal, ESC are able to divide with daughter cells retaining an undifferentiated phenotype. This trait has been shown to continue without limit, with ESC undergoing many cellular passages and still remaining in an undifferentiated state [2]. However, these characteristics are not guaranteed *in vitro*, and careful culture conditions must be maintained to sustain this self-renewing state. These conditions include soluble cues in the media as well as the substrate to which the ESC attach [12]. The first

successful culture configuration which sustained hESC self-renewal included the cells being seeded on a mouse embryonic fibroblast (MEF) feeder layer [2]. Further culture configurations have since been developed, including a feeder-free system, in which the hESC are seeded onto Matrigel^{TM}, a gelatinous protein mixture consisting of various extracellular matrix (ECM) components, with MEF conditioned media [13]. These special culture conditions have been shown to sustain pluripotency and self-renewal for an extended period of time.

In addition to retaining the self-renewal state, modulation of the extracellular microenvironment is necessary to direct differentiation to specific lineages (Figure 1.2). Through this modulation, the feasibility of differentiation of ESC to most of the tissue types in the body has already been established [14-16]. Environmental manipulations to direct differentiation are often achieved by addition of soluble factors to the culture media. For example, D'Amour *et al*. uses a series of different conditioned media to mimic *in vivo* pancreatic development for endocrine cell production *in vitro*[17]. While soluble cues are predominantly used to modulate stem cell fate, more recently cues from the underlying substrate have also been shown to have a significant role in stem cell fate commitment. Substrates are widely known to be needed as an anchor for most cell types *in vitro* and *in vivo*; in addition, they also can be used as a means to induce and guide differentiation. Numerous aspects of the stem cell niche can influence the cues from the substrate, including extracellular matrix (ECM) and substrate topology (reviewed in [18, 19]). Crucial cues from the underlying substrate which are increasingly gaining importance, particularly in the area of stem cells, are the physical properties of the substrate, in particular, substrate mechanical properties [20].

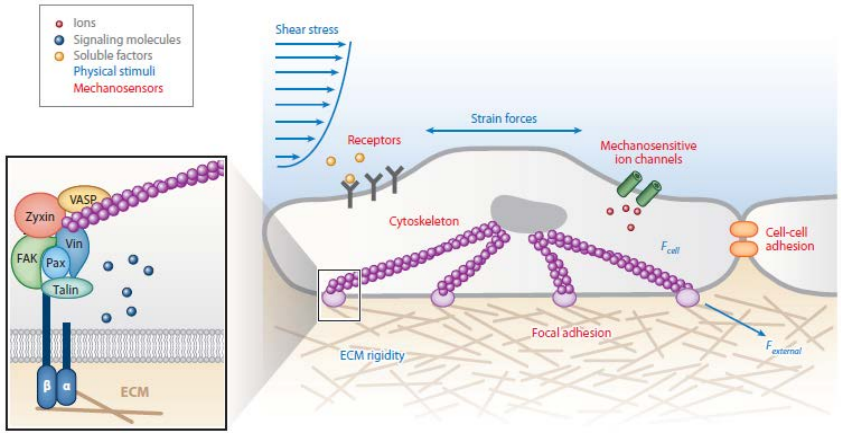

Figure 1.2. Schematic of factors affecting cellular behavior

Soluble cues, such as ligands binding to receptors, and insoluble cues, such as ECM and cell-cell contact, act to guide cellular behavior through effected signaling pathways [19]

Therefore, the primary characteristics of ESC are that they can proliferate indefinitely in an undifferentiated state and can differentiate to tissue specific lineages *in vitro* [1, 2]. Both of these processes have gained considerable attention within the scientific community, and much research has been invested in understanding and improving these processes so they may be used for regenerative therapy. Propagation of self-renewing cells is needed for scale-up of undifferentiated cells, and differentiation is needed to guide these cells to the tissue of interest. Improving these processes and making them more efficient has, for the most part, been purely experimental. Studies have focused on developing cause-and-effect relationships by perturbing soluble and insoluble cues and experimentally quantifying the cellular effect, which can be costly and time consuming. While mechanistic information can improve on this process, obtaining this information from the complex ESC system using a purely experimental approach is often difficult. Mathematical models are invaluable in gaining this mechanistic insight of biological systems.

1.3 USEFULNESS OF MATHEMATICAL MODELS IN ANALYZING STEM CELL DIFFERENTIATION

Mathematical models have been extremely successful in understanding and analyzing biological and cellular systems. These approaches have been less explored in the ESC system, and a more dedicated effort is needed to extract the full potential of mathematical models applied to ESC. However, initial efforts by a few prominent groups have demonstrated the promise of mathematical approaches in extracting mechanistic information from stem cell systems.

A key aspect of ESC which governs their self-renewal capabilities is the gene regulatory network. Several key genes have been identified which govern pluripotency, and include Oct4, Nanog, and Sox2 [21-24]. Understanding how these genes interact can give insight into self-renewal behavior, and mathematical models have been utilized to gain a more quantitative understanding of the system. Notable studies of this network include the study by Chickarmane *et al*. [25] which reports identification of a bistable switch in the Oct4-Sox2-Nanog network leading to a binary decision of the cells to self-renew or differentiate (Figure 1.3). In a follow up work [26] the authors further extend the model to incorporate lineage specific differentiation namely to endoderm and trophectoderm. MacArthur *et al.* [27] also analyzed the Oct4-Sox2-Nanog network coupled with a lineage specification network to investigate the induction of pluripotent cells from somatic cells. Glauche *et al.* also look at noise associated with this gene circuit, but with regards to pluripotency [28].

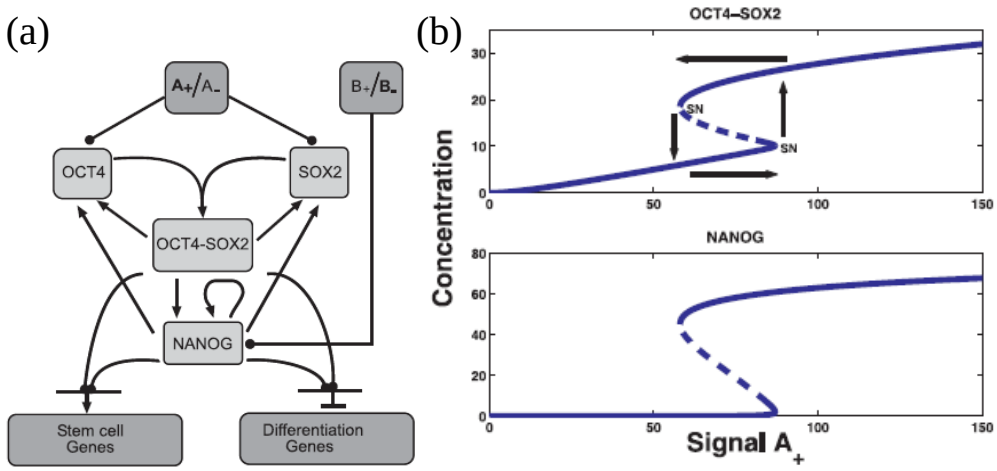

Figure 1.3. Pluripotency control model

(a). Transcription factor network and connections governing pluripotency. (b) Resulting bi-stable switch in protein levels with signal B as proposed by Chickarmane *et al*. [25]

In addition to gene regulatory networks, mathematical approaches have also been utilized to understand signaling pathways involved in pluripotency and differentiation. Prudhomme *et al*. [29] performed a thorough systematic analysis of how the intracellular signaling relates to different extracellular cues during differentiation of mouse ESC. A partial-least-squared multivariate model was built to show the role of signaling proteins in self-renewal, differentiation, and proliferation of stem cells. In a follow up work, Woolf *et al*. [30] investigated the signaling network to determine the "cue-signal-response" interactions through a Bayesian network algorithm. The nodes of the network are assigned to be an extracellular stimulus, a signaling protein, or a cell response, following which the model identified interconnections between nodes without being explicit about the nature of connections (inhibition, induction). Mahdavi *et al.* [31] focused on signaling networks as well, employing sensitivity analysis in the Stat3 pathway to predict how self-renewal in mouse ESC is controlled.

Beyond intracellular processes, there have been modeling efforts to describe population behavior during self-renewal and pluripotency, and how this behavior is manifested from single cell characteristics. Viswanathan *et al.*[32] proposed a single cell model of the ESC system which would account for the heterogeneity in the cell population (Figure 1.4). They based their model on number of ligands/receptors per cell, and predicted the behavior of ESC self-renewal and differentiation, and the system's response to different exogenous stimuli. Such analysis has potential use in selecting specific tuning parameters while guiding ESC towards a specific fate [32, 33]. Prudhomme *et al*. [34] developed an ordinary differential equation based kinetic model to quantify the differentiation dynamics in response to combinations of different extracellular stimuli. Based on experimental data of ESC response to different combinations of extracellular matrix and cytokines, the authors estimated kinetic rate constants for each culture condition.

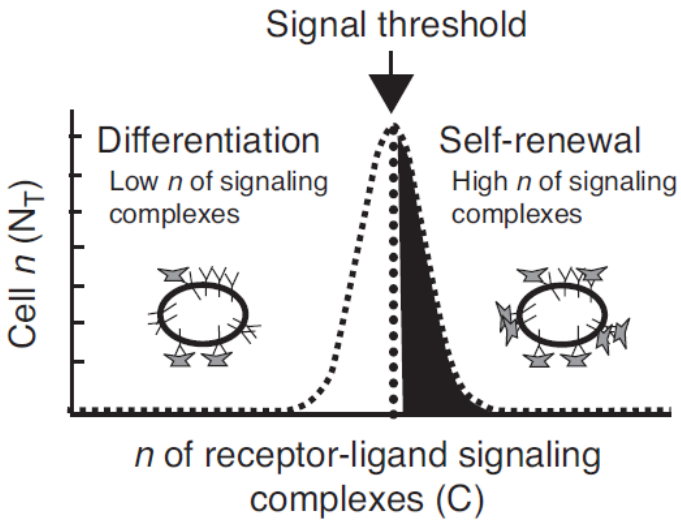

Figure 1.4. Single cell ESC model

Receptor-ligand population distribution as proposed by Viswanathan *et al.* [32]

ESC are heterogeneous in nature, with high variability of mRNA and protein expression within a population, both during self-renewal and differentiation. Therefore, mathematical analyses that assume homogeneous populations may not be sufficient to describe ESC population dynamics, and therefore more descriptive techniques are needed to model population heterogeneity. A common modeling technique to describe population dynamics in a heterogeneous system is the population balance equation (PBE), which has been used to model various systems, including adult stem cell behavior [35]. Other approaches to capture the dynamics of a heterogeneous cellular population have also been developed. One notable example is the cellular ensemble model, in which individual cells are tracked with time, with the behavior of the individual cells dictated by rules or equations which are solved for each cell in the population [36]. Distributions and variability associated with the parameters of these rules and equations can capture the heterogeneity in the population. This approach has been successfully used in the hematopoietic system. Glauche *et al*. utilized this model to describe lineage specification of hematopoietic stem cells, with cellular choices governed by a competition of different lineage propensities [37]. Through their model, various cellular populations are tracked with time, and insight into the differentiation process is obtained.

Another prominent feature of ESC, both during self-renewal and differentiation, is the cell cycle. ESC have a short doubling time, mainly due to an abbreviated G1 phase [38, 39]. In contrast, somatic cells have a much longer doubling time with the majority of the cell cycle spent in the G1 phase. When ESC differentiate, this G1 phase elongates, resulting in an overall longer doubling time [40, 41] and slower propagation. Analysis of the cell cycle behavior of ESC has largely remained experimental to date. However, mathematical modeling in other systems has demonstrated the usefulness of this approach in gaining mechanistic insight of cell cycle

behavior. Much focus has gone into describing in a mathematical sense how proteins governing the cell cycle work to control phase transitions [42, 43] (Figure 1.5). Understanding these control mechanisms has been invaluable to explain phenomena which is not intuitive from experimentation, and can offer insight into wide variety of observed behaviors, including sensitivity [44], reversibility and irreversibility [45, 46], and cancer behavior [47]. In particular, much work has gone into describing the G1-S transition and associated restriction point [48, 49]. Through mathematical modeling, it is now a current belief that this checkpoint is governed by a bistable switch generated by protein feedback networks [49, 50] , something which has been validated experimentally [51].

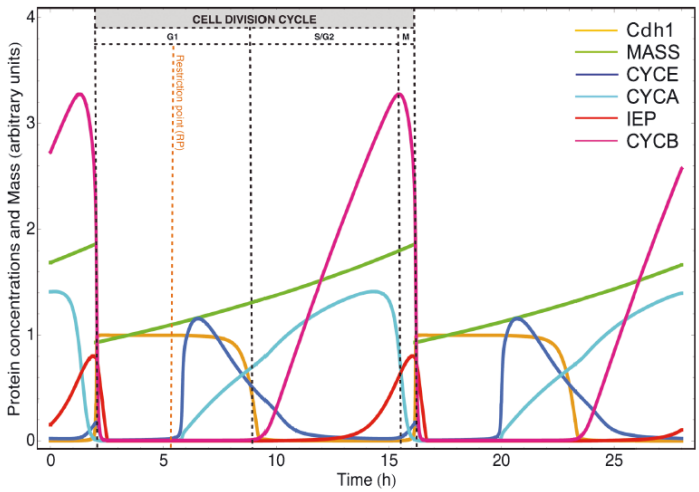

Figure 1.5. Simulated protein dynamics and relationship to cell cycle

ODE model captures temporal trends of cell cycle indicative proteins and how they change with phases [49]

Mathematical descriptions of cell cycle population dynamics have also received attention [52, 53]. These models do not focus so much on intracellular events, but describe how populations of cells in the individual cycle phases change with time and environment. A major

application of this type of modeling is in cancer research, for which it is often necessary to know cell cycle phase behavior to guide treatment and to track cellular growth [54-58]. Another application is to describe synchrony behavior [59-61], a major factor of which is cell cycle variability.

1.4 SPECIFIC AIMS

Our long-term goal is to pave the way for the use of stem cells for cell replacement therapy for diabetes. The last decade of intense research in this area has clearly highlighted the need for more in-depth mechanistic understanding. Mechanistic understanding of complex systems like stem cells is best mediated by experimentally informed mathematical models, which is the objective of the current project. Stem cell differentiation is induced by controlled manipulation of the cell microenvironment. Stem cells transmit this information to the nuclei which activates specific gene regulatory networks governing differentiation. While this is a single cell-level phenomenon, the entire population responds to these external cues with certain variability and heterogeneity. The objective of this project is to characterize the stem cells at these levels: (i) intracellular information, specifically regulatory network identification; (ii) extracellular environment, and the influence of substrate characteristics on differentiation; and (iii) population information from differentiation and cell cycle dynamics. At each of these levels, the analysis of variability and complexity deserves particular attention. We individually address each of these cellular levels by the following specific aims:

1.4.1 Specific Aim 1: robust identification of gene regulatory networks in the presence of intracellular noise

Networks of gene interactions are the primary determinant of cell fate and differentiation. Determining gene network interactions from experimental gene expression data is a critical, yet challenging, task. The variability in the mRNA levels further enhances the difficulty of this exercise. We developed a novel optimization formulism to identify robust gene interactions from noisy gene expression dynamics, which is detailed in Chapter 2.

1.4.2 Specific Aim 2: identification of specific attributes of extracellular substrates influencing ESC differentiation

Most cells in the body require an associated substrate with which they can attach and remain functional. In addition to promoting viability, substrates also guide cellular behavior. While soluble cues are the most predominant environmental perturbation used to drive stem cell differentiation, manipulation of insoluble biophysical cues, including cellular substrates, have also been shown to have a significant influence on cell fate determination. However, deciphering the mechanisms by which biophysical cues affect cells becomes challenging due to the often complex nature of the substrates. Because of this complexity, analysis of the cause-effect relationship between cells and their associated substrates is often not possible with a purely experimental approach. We utilized a system level approach to correlate the different factors associated with natural substrates with differentiation of ESC, thereby gaining insight on the importance of microenvironment on the stem cell system. This approach is detailed in Chapter 3.

1.4.3 Specific Aim 3: analyzing population dynamics of ESC during self-renewal and differentiation

While the mechanism of self-renewal and differentiation can be assumed to be identical in all cells, the magnitude of expression of gene or protein typically varies across the population. Hence, even though we may assume a cell population to be 'differentiated,' the level of differentiation, as judged by expression of specific proteins, typically varies across the population. We developed a stochastic population model to capture this heterogeneity of the ESC system and elucidate mechanistic information on ESC processes, primarily focusing on two processes: cell cycle and initial lineage commitment. In Chapter 4 we discuss the development and results of a stochastic population model to capture the cell cycle behavior both during pluripotency and differentiation. This model was extended to describe initial differentiation during definitive endoderm induction (Chapter 5). We show in these two chapters that utilizing these population models allows for the identification of plausible differentiation and cell cycle mechanisms in the presence of population heterogeneity.

2.0 ROBUST IDENTIFICATION OF GENE REGULATORY NETWORKS IN THE PRESENCE OF INTRACELLULAR NOISE

2.1 INTRODUCTION

In the first aim of this work, we will consider intracellular gene regulatory networks, specifically reverse engineering the networks from time series gene expression data. Gene regulatory network identification is an important problem, and yet accurate inference of gene interactions is made particularly difficult due to the inherent noise in transcription. Network identification algorithms therefore require numerous experimental replicates for reliable conclusions because of this inherent noise. Furthermore, evidence of robust algorithms directly exploiting basic biological traits are few. Advanced gene network identification algorithms are therefore expected to be efficient in their performance and robust in their prediction. This chapter presents a novel formulism to robustly identify gene regulatory networks in the presence of biological noise [62].

Biological systems have been shown to be “robust-yet-fragile” [63]. A robust system property is insensitive to a set of system perturbations. In the cellular system, perturbations could include changes in the extracellular environment or stochastic fluctuations in intracellular protein concentrations. In contrast, the same system can be fragile, in that other perturbations can cause devastating effects on the organism. The nature of the network governing the system largely

dictates its robustness and fragility, and can be categorized into two main structures: sparse and redundant. Network sparsity refers to a system with as minimal connections between nodes as possible to perform a specific task; for the specific case of gene regulatory networks, sparsity refers to the observation that interactions between transcription factors guiding gene expression are minimal. This is in direct contrast to network complexity and redundancy. Redundant systems are those which have components and connections which perform the same, or similar, tasks [64]. This is to increase reliability, and to ensure that the goal of the network is accomplished in the face of system perturbations. While redundancy is prominent in numerous cellular processes, including metabolic networks [65, 66], and complex network connections have been previously reported to result in robustness, new evidence has demonstrated that fewer connections are favorable in gene networks due to the costs associated with dense systems [67]. Network sparsity has been experimentally observed in various gene networks, including those of E. Coli, yeast, and sea urchin [67-70]. In this light, the focal point of the identification algorithm is network sparsity. Governed by the hypothesis of sparsity of gene network connections, the target of our network identification algorithm is to find the network structure with minimum number of connections that is in agreement with the experimental data at an acceptable level of tolerance. We initially developed an optimization formulation to identify the regulatory network from time profiles of gene expression data [71]. This algorithm was based on the following approximations: (i) gene expression dynamics were approximated by linear ordinary differential equations (ODE); and (ii) the system was treated as deterministic by considering only the mean experimental data for the analysis. We further developed this algorithm to utilize bootstrapping to identify robust networks from noisy data. The aforementioned approximations are removed by (i) representing the gene expression profile with an S-system model and (ii) directly accounting

for variability in experimental data. Our algorithm enables identification of robust networks from an inherently nonlinear and noisy system. We test the performance of our algorithm in various case studies including *in silico* and experimental data sets.

2.2 METHODS

2.2.1 S-System representation of gene expression dynamics

Identification of the regulatory network from time series gene expression data first requires modeling the dynamic evolution of the individual genes constituting the network. Here we model gene dynamics as a set of coupled nonlinear ODE following the S-system formulation, which captures the nonlinearity in gene expression profiles using a power-law kinetic representation.

For a system with N-genes, the S-system model can be represented using Equation 2.1:

$$\dot{X}_i = \alpha_i \prod_{j=1}^{n} X_j^{g_{ij}} - \beta_i \prod_{j=1}^{n} X_j^{h_{ij}} \qquad (2.1)$$

Where X_i is the concentration of the gene *i,* α and β represent the kinetic rate constants, *g* and *h* represent the kinetic orders for the production and degradation terms, respectively, and *n* is the total number of species in the system, in this case the total number of genes in the network. In this work, we are using a modification of the above equation by assuming that species degradation follows first order kinetics of the corresponding species and independent of other

species (h_{ij}= 1 for $i=j$; 0 otherwise). While being relevant to biological systems [72], this assumption also reduces the unknown parameters from $2n(n+1)$ to $n(n+2)$ [73].

2.2.2 Network identification algorithm

Our network identification algorithm is primarily based on the hypothesis of sparsity of network connections governing gene transcription. Hence, our overall objective is to determine the sparsest gene regulatory network which can satisfactorily capture the observed network dynamics. Following this idea, the network identification problem is formulated as an optimization problem with the objective of promoting sparsity given the constraint of maximizing predictive capacity. Such a problem definition results in a bi-level optimization problem, where the constraint itself is an unconstrained optimization problem. In the current formulation using S-system to model the gene expression level (Equation 2.1), the kinetic orders (g_{ij}) are decomposed into two parts: binary part, λ_{ij}, which determines the existence of the connection; and continuous part, ρ_{ij}, representing the nature and strength of interaction for an existing connection. A value of 1 of the binary variable λ_{ij} would indicate the presence of the corresponding connection $X_i \leftarrow X_j$, while value of 0 indicates its absence. These binary variables are optimized in the upper level which results in an integer programming problem. For each chosen network in the upper level, the connections are sent to the lower level, where corresponding ρ_{ij} are optimized to maximize network prediction and hence minimize deviation of the network predictions from the observables. The lower level essentially optimizes both strength (magnitude) and nature (sign) of the existing connections (ρ_{ij} , reactions orders) as well as the strengths of the production and degradation rate constants (α_i, and β_i respectively). Hence it

results in a continuous nonlinear programming problem where the objective is to minimize the deviation of the predicted profiles from experimental data in a least square sense. A constraint of tolerance (*tol*) is imposed on this minimized error which defines the maximum allowable deviation in prediction. The mathematical formulation of the network identification problem in its entirety is shown in Equation 2.2:

$$\begin{aligned} &\min_{\lambda} \quad J = \sum_{i,j=1}^{n} \lambda_{ij} \\ &S.T.: \; C_1 = \arg\min \chi(\lambda) \le tol \\ &\qquad C_2 = 1 \le \sum_{i,j=1}^{n} \lambda_{ij} < n \times (m-3) \end{aligned} \tag{2.2}$$

where

$$\min \quad \chi = \left[\sum_{t=1}^{nstep} \sum_{i=1}^{n} \left(x_{t,i}^{\exp} - x_{t,i}^{pred} \right)^2 \right]^{\frac{1}{2}}$$

$$\alpha, \beta, \rho$$

where

$$\frac{dx_i}{dt} = \alpha_i \prod_{j=1}^{n} x_{i,j}^{g_{ij}} - \beta_i \prod_{j=1}^{n} x_{i,j}^{h_{ij}}$$

$$g_{ij} = \lambda_{ij} \cdot \rho_{ij}$$

$$h_{ij} = \begin{cases} 1, & i = j \\ 0, & otherwise \end{cases}$$

$\lambda_{i\,j}$= binary variable

$x_{ij}^{\exp}, x_{ij}^{pred}$ = experimental and predicted gene expression levels, respectively

α_i, β_i= kinetic rates of gene *i* production and degradation respectively

g_{ij}, h_{ij} = kinetic orders of the effect of gene j on the production and degradation of gene i, respectively

nstep = number of time points

n = number of genes constituting the network

m = number of experimental time points

In the above formulation $\sum\lambda$ represents the total number of network connections, minimizing which will promote sparsity in the network. The upper level integer programming is solved using combinatorial optimization techniques since combinatorial approach is known to handle L_0 minimization problems more efficiently than approximation algorithms [74]. Of them, evolutionary algorithms are particularly efficient in finding a good approximate solution for combinatorial problems [75]. In this work, we have used genetic algorithm (GA) for solving the integer programming problem, while the lower level nonlinear programming problem is solved using a standard least square optimization routine.

GA is typically designed to handle unconstrained optimization problems. One technique for constraint handling in GA is by a penalty function, where the constraint is conditionally incorporated in the objective function. For conditions violating the constraint the objective function is penalized, and not so otherwise. In the current formulation the constraint is incorporated in the objective function using the following modification to the objective function:

$$\phi = \min\left[\sum_{i,j=1}^{n}\lambda_{ij} + penalty * \frac{\max[\zeta,0]}{\zeta}\right] \qquad (2.3)$$

$$\text{where } \zeta = \frac{\arg\min \chi(\lambda)}{\boldsymbol{tol}} - 1$$

A significant advantage of the bi-level formulation is that it allows optimum utilization of experimental data by sequentially reducing the number of unknown parameters in the lower level. In a conventional least-square parameter estimation problem, the connectivity is fixed and includes all possible network connections. Therefore, the size of the identifiable system is restricted, governed by the availability of experimental data points so that the number of unknown parameters is less than the number of data points. For instance, a single level algorithm, using the above S-System formulation, would be restricted to less than m-3 genes. However, in the current bi-level formulation, this restriction is relaxed. Because the number of network connections is first reduced in the upper level, the number of genes to be analyzed is not so restricted, with the only constraint coming from the connectivity:

$$\sum_{i,j=1}^{n} \lambda_{ij} < n \times (m-3) \tag{2.4}$$

Hence the constraint is imposed on the maximum number of binary variables assigned in the upper level, but does not constrain the total size of the analyzed network. Moreover, our primary objective being sparsity of network connections, the formulation essentially tries to minimize the number of connections assigned to 1. Hence, except for the very initial phase of GA evolution the constraint defined in Equation 2.2 typically does not become active, and never so in the final optimal solution.

2.2.3 Identification of robust networks

Real world data typically contains noise due to experimental uncertainty and system stochasticity. Biological data are particularly notorious for its inherent heterogeneity and stochasticity [76]. Hence, it is important to explicitly account for data variability in order to increase confidence in the predicted network. In the presence of large experimental repeats it may be possible to determine robustness of the identified networks by repeatedly solving the network identification problem at each of the experimental data sets and analyzing the connections which are heavily repeated. However, drawing statistically significant inference would necessitate a large data set which is often impractical to obtain with gene expression dynamics.

An alternative to actual experimental repeats is to use bootstrapping. The purpose of this statistical technique is to estimate the distribution of the estimator $\hat{\theta}$ around the unknown true value θ. However, instead of achieving this with a large number of individual replicates, bootstrapping utilizes resampling of the data. In this way, a large number artificial data sets can be generated from a limited number of experimental repeats. For each bootstrap run, data samples are randomly chosen, with replacement, from the empirical distribution, with the size of each artificial set being the same as the experimental set (e.g. if the experimental set has 20 data points, so would the bootstrap set). For each bootstrap, the estimators (e.g. mean, variance, or, as in the case of the current work, regression parameters) are calculated, and with sufficient number of resampled data sets, relevant statistical information, including confidence intervals, can be estimated [77, 78].

In our algorithm, we are dealing with limited experimental data. Hence, following the above methodology, we generate a large artificial dataset by repeated resampling of the limited experimental repeats. Once the bootstrapped samples are obtained, the network identification algorithm described in section 2.2.2 is applied to all bootstrap data sets to identify a network corresponding to each. The network sets thus obtained are further analyzed to determine the frequency of occurrence of each connection in the entire set of identified networks. We hypothesize that frequent occurrence of network connections in the bootstrap samples indicate the insensitivity of the corresponding network to experimental noise, and hence claim that connection to be robust.

In order to quantify the quality of prediction of the proposed algorithm the measures of *recall* and *precision* are used, calculated as:

$$\begin{aligned} recall &= \frac{TP}{TP + FN} \\ precision &= \frac{TP}{TP + FP} \end{aligned} \tag{2.5}$$

Where: TP (True Positive) denotes the number of connections correctly captured; FN (False Negative) denotes existing connections which are not captured in the identified network; and FP (False Positive) denotes connections which are incorrectly captured in the identified network. Following the above equation, a low value of recall would indicate a more conservative estimate which is unable to capture many of the existing connections; a low value of precision will indicate prediction of incorrect connections not appearing in the actual network; and a value of 1 will indicate perfect network identification. The flow diagram of the overall network identification algorithm is shown in Figure 2.1.

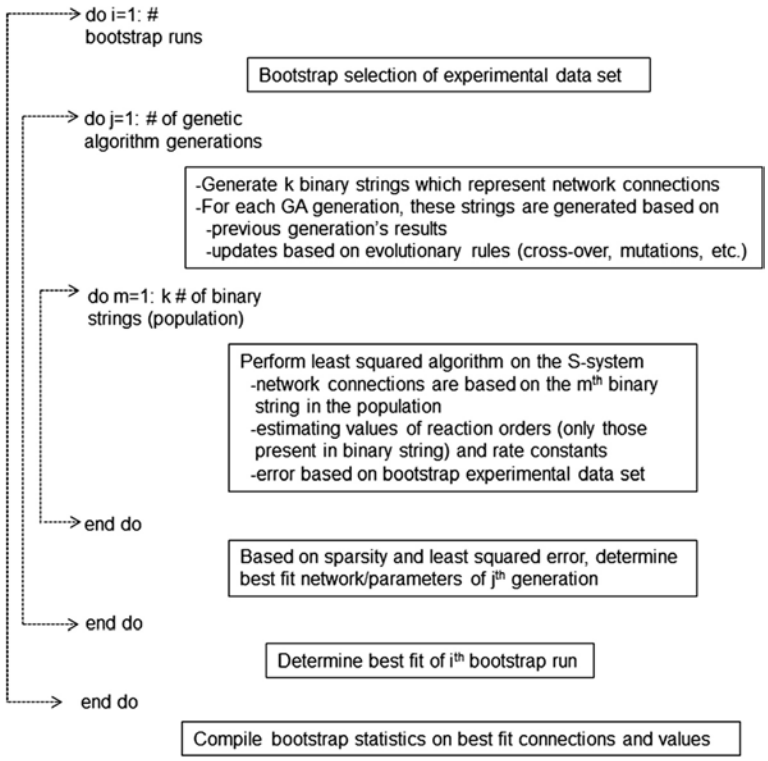

Figure 2.1. Pseudo-code of the robust network identification algorithm implementation

2.3 RESULTS

The performance of the developed bi-level integer programming algorithm is demonstrated on three case studies. In the first case study, we consider *in silico* gene expression data generated from a benchmark artificial 5-gene network model. In the second case study, the applicability of the algorithm on a larger network is tested using an *in silico* 10-gene network. In

the third case study, the algorithm is applied to an experimental data set of the SOS DNA repair system in *E.coli.*

2.3.1 Case study 1: five gene network model

The purpose of this case study is to validate the algorithm on a small network with and without experimental noise. The chosen 5-gene network model [73] has been used as a benchmark problem by different research groups to test the validity of their algorithms [79, 80].

2.3.1.1 Network identification without noise

Using the S-system formulation, the 5-gene network model can be represented by the system of five coupled nonlinear ODE, shown in Equation 2.6 [73].

$$\begin{aligned}
\dot{X}_1 &= 5X_3X_5^{-1} - 10X_1 \\
\dot{X}_2 &= 10X_1^2 - 10X_2 \\
\dot{X}_3 &= 10X_2^{-1} - 10X_3 \\
\dot{X}_4 &= 8X_3^2X_5^{-1} - 10X_4 \\
\dot{X}_5 &= 10X_4^2 - 10X_5
\end{aligned} \tag{2.6}$$

In order to test our identification algorithm on this model we first generate *in silico* data by integrating these equations, which we use as experimental data for the identification algorithm. The limits of integration used were t={0, 0.5} (a.u.), with initial conditions 0.1, 0.7, 0.7, 0.16, and 0.18 for the five genes, respectively. For each gene, twenty time points of the temporal profile were generated through the numerical integration, yielding a total of 100 data

points (excluding initial conditions) for the system. To formulate the bi-level optimization problem, $n^2 = 25$ binary variables are introduced corresponding to each of the five connections. GA, used to solve the upper level integer programming problem, does not have a convergence criterion. Standard practice is to evolve the population for enough generations until no significant improvement is observed. While dependent on the system being optimized, our experience has shown that if there is no change in the fitness value after 50 generations, there will unlikely be any further change henceforth. Figure 2.2 (a) illustrates the convergence characteristics of the GA for this example; the algorithm was run for a total of 150 generations, although at over 103 generations, the optimal output remained invariant. The efficiency of the algorithm depends on appropriate choice of starting population, as well as other involved parameters, in addition to the number of generations. The initial population size plays an important role in the quality and efficiency of the algorithm. A small population size may lead to local convergence or extremely large number of generations. To avoid that a population size of 20 was chosen and the algorithm evolved for 150 generations. The crossover probability is chosen to be at a standard value of 0.5, and the chosen mutation probability of 0.02 was expected to maintain diversity in population. Since the data contain no noise, the tolerance in the lower level least square optimization problem has been kept at a very low value (10^{-5}). Typically least square optimization routines are very sensitive to the user defined initial guess. To make sure that the algorithm can identify the underlying network structure even without any *a priori* information, we deliberately assigned the initial guess values for the least square optimization problem to be largely different from the actual values, and tested the algorithm for various combinations of the initial guess.

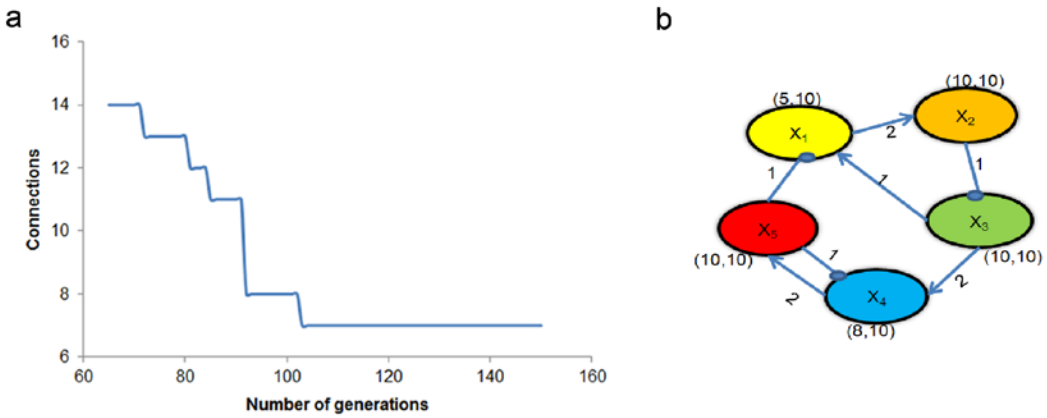

Figure 2.2. Identification of a 5-gene network without noise

(a) Convergence study of the genetic algorithm. The number of connections identified in each of the solutions generated by GA is plotted. No feasible solution was found with less than 65 generations. (b) Identified network. Arrows represent the positive regulation and the filled circles represent the negative regulation of the genes. Kinetic orders of each connection are represented above the corresponding connecting lines and the rate constants for each gene are shown above the genes. All connections and parameters are consistent with the original differential equations used to generate the *in silico* data.

Figure 2 .2(b) illustrates the 5-gene network identified using the above formulation. The kinetic orders (g_{ij}) are depicted over the connection and the kinetic rate constants (α_{ij}, β_{ij}) are depicted in brackets. The precision and recall value were both a perfect 1.0, indicating the accuracy with which the proposed algorithm predicted the network structure from time profile gene expression data. In addition, the identified kinetic orders and rate constants are also in agreement with the actual network model presented in Equation 2.6. These results validate the performance of the algorithm for a small network under deterministic conditions.

2.3.1.2 Network identification under data uncertainty

The performance of the algorithm is next analyzed in the presence of experimental noise, generated by adding 5% Gaussian noise to the time-course data generated from Equation 2.6.

The number of samples in each time series remained the same as in the study without noise: 20 data points per gene. Therefore, for each gene, at each of the 20 time points, three new data points were generated by adding 5% Gaussian noise to the data point; these three new points represent three experimental replicates of the samples. These three data sets are then resampled using bootstrapping to generate 1000 artificial data sets. The network identification algorithm was then applied at each of the data sets to generate 1000 alternate networks. The presence of noise in the data restricts the accuracy by which the predicted profile can agree with the data. Hence the tolerance, *tol* in Equation 2.2, was relaxed to 0.12. If higher levels of noise are suspected, this tolerance level would have to be increased. The GA code was evolved for 200 generations while retaining the population size of 20. The ensemble of alternate networks thus generated was analyzed for frequency of appearance of each of the connections (Figure 2.3(a)) which was hypothesized to directly correspond to its robustness against experimental noise.

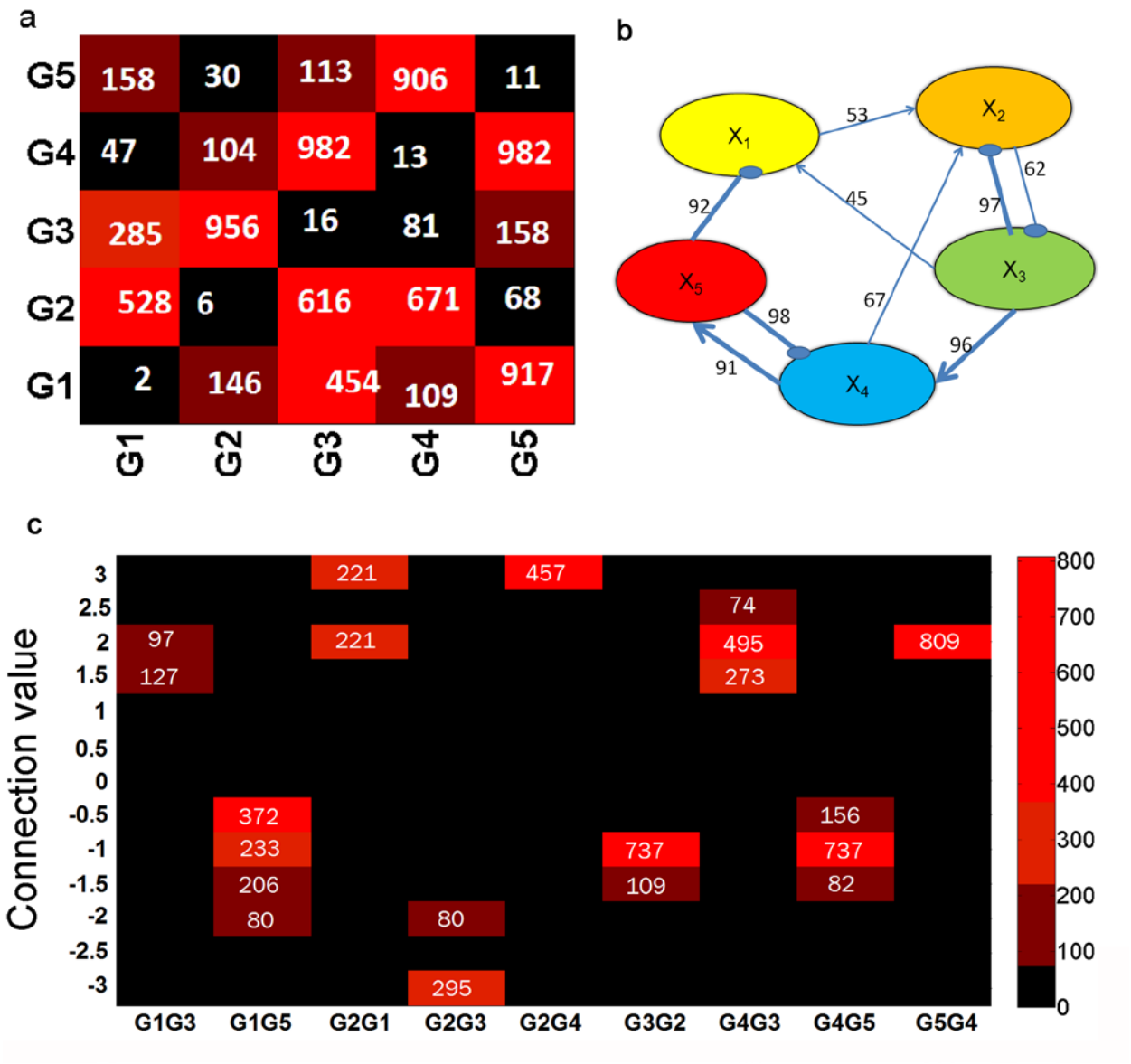

Figure 2.3. Results from 5-gene network identified under data uncertainty with 5% noise

(a) Number of bootstrap occurrences for each connection (1000 bootstrap samples total). (b) Identified network structure. Numbers above each connection represent percent occurrence, with the thick lines representing the number of connections appearing in more than 90% of the bootstrapped samples and the thin lines representing the connections appearing in more than 45% of bootstrapped samples. (c) frequency of specific connection values shown as heat map

Figure 2.3 (b) further illustrates the identified robust network connections screened for 45% occurrence, with frequency of occurrence of network connections being depicted over the connection. Quite encouragingly, the algorithm correctly identified all the existing connections in the actual network. However because of noise, the algorithm also identifies two false interactions involving gene 2, hence resulting in a recall and precision of 1 and 0.78, respectively.

The expected values of the S-system parameters estimated at 90% confidence level are represented in Table 2.1(a) (g_{ij}) and Table 2.1(b) (α_{ij}, β_{ij}), which demonstrates the excellent performance of the algorithm in identifying network parameters even from noisy data. The error of the rate constants varies considerably, from 1 to 80%. This is due to the insensitivity of the network identification to these parameters. This is in contrast to the sensitivity of the identification to the reaction orders; the errors on these parameters are low (0-20%) ,demonstrating the accuracy of the network identification. The heat map in Figure 2.3 (c) further shows the algorithm's effectiveness in finding a tight range of reaction orders of the robust connections in the network.

Table 2.1. Comparison of predicted and actual values of the S-system parameters for 5-gene network

(a) Reaction orders. (b) Rate constants.

a

Connection	g_{actual}	$g_{estimated}$
G1G3	1	1.2±0.09
G1G5	-1	-1.0±0.03
G2G1	2	2.4±0.04
G2G3	NA	-3.4±0.06
G2G4	NA	3.9±0.05
G3G2	-1	-1.1±0.03
G4G3	2	1.9±0.02
G4G5	-1	-1.0±0.01
G5G4	2	2.0±0.02

b

Gene	α_i		β_i	
	actual	estimated	actual	estimated
X1	5	3.8±0.2	10	18.0±0.8
X2	10	13.8±0.9	10	16.2±0.2
X3	10	13.8±0.2	10	11.2±0.23
X4	8	8.1±0.1	10	11.8±0.1
X5	10	10.3±0.05	10	8.9±0.03

To evaluate the accuracy of the formulism under increased uncertainty, the algorithm was tested under various amounts of added noise. As one would expect, the accuracy of the algorithm depends on the level of noise added to the *in silico* data. Table 2.2 shows this trend, with the precision and recall being compared with 5, 7, and 10% noise. Increasing noise increases the number of false negatives, thereby reducing the recall. Interestingly, precision actually improves with increasing noise, indicating less false positives. This trend seems to converge, with both the recall and precision holding constant at 7 and 10%. Figure 2.4(a) shows the identified network with a data set incorporating 10% white Gaussian noise. The algorithm does not identify any connection which is not in the actual network (e.g. 0 false positives) and is

therefore able to achieve a perfect precision. However because of noise, the algorithm also fails to identify three of the actual connections (false negatives), hence resulting in a recall of 0.57. To be considered robust, a connection needs to be present in a certain fraction of bootstrapped identified networks. This fraction, called the bootstrap threshold value, affects the recall and precision of the algorithm, as shown in Figure 2.4(b). Given the identification sensitivity to this threshold, the value should be chosen judiciously depending on the overall goal. For instance, if the objective is to identify all connections without being concerned with false connections, recall should be high, and therefore a lower bootstrap threshold should be chosen. Conversely, if the priority is to avoid identification of false connections, then precision should be high, and a higher bootstrap threshold should be chosen.

Table 2.2. Effect of added noise on the network identification results

Percent noise	Recall	Precision
5	1	0.78
7	0.57	1
10	0.57	1

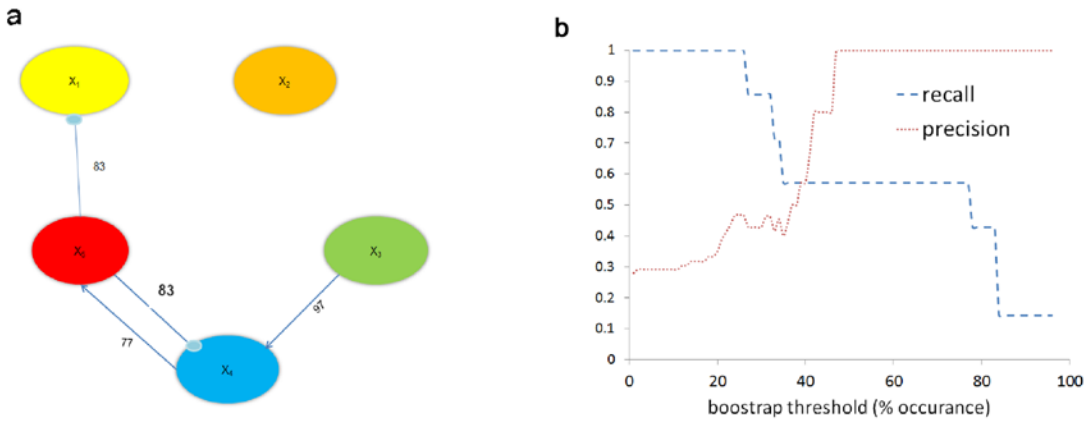

Figure 2.4. Results from 5-gene network identified under data uncertainty with 10% noise

(a) Identified network structure. Numbers above each connection represent percent occurrence. (b) Sensitivity of recall and precision to bootstrap occurrence threshold.

While the analysis is performed on 1000 bootstrap samples, it is computationally expensive to solve 1000 network identification problems. Hence, we investigated the sensitivity of the identified robust network on the number of bootstrap samples by considering a broad range of samples from 200 to 1000. Figure 2.5 illustrates the percentage of total number of appearances of each identified interaction in every 200 bootstrapped samples, using 5% noise. The difference in the maximum and minimum number of appearances is less than 8% for all connections. The clearly shows that as little as 200 bootstrap samples can be enough in drawing statistically significant conclusions, which is in agreement with the literature [77].

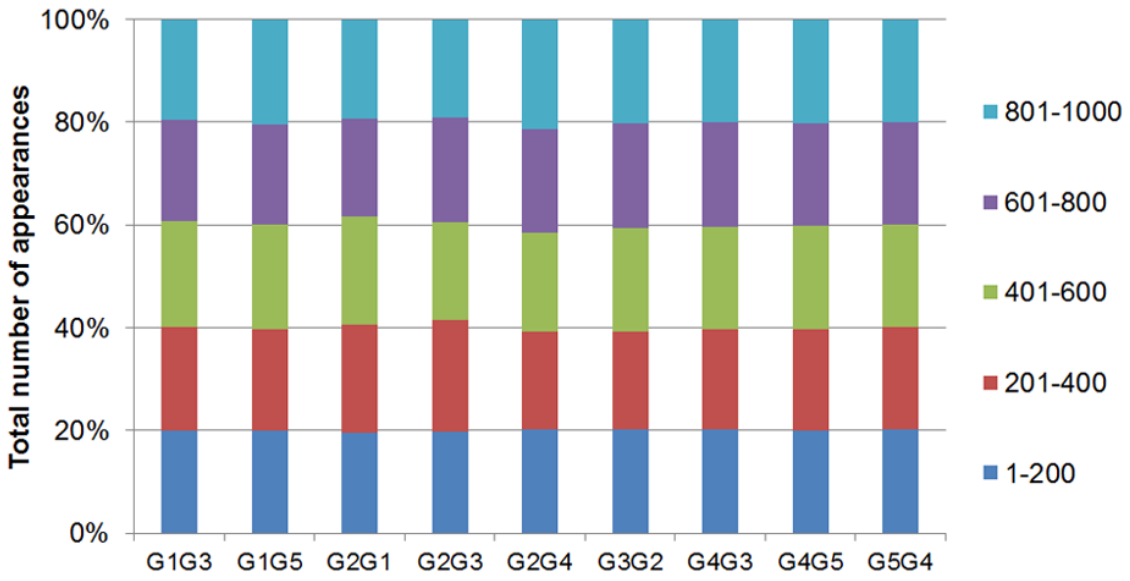

Figure 2.5. Convergence study on network identification using bootstrapping with 5% noise

2.3.1.3 Deterministic network identification under data uncertainty

To assess the necessity of this bootstrapping technique, the aforementioned results were compared to a control group which did not utilize bootstrapping. To do this, a more deterministic approach was employed. Experimental replicates were generated as detailed: 10% white Gaussian noise was added to the 5-gene *in silico* network. Instead of bootstrapping these replicates, the deterministic network identification was performed on the mean of the replicates. This was done for 3, 5, 7, 9, 12, 15, 18 and 20 replicates with the resulting precision and recall calculated for each case; results are shown in Figure 2.6. As shown, when the input data is generated from fewer than seven replicates, a solution is not found. Even with seven replicates, the results are relatively poor. While the recall is comparable to that generated from bootstrapping (~0.57), precision is much worse (0.5). As the number of replicates is increased, this precision increases; however, even at 20 replicates, precision is not perfect (0.8).

Furthermore, in practice, generating this many experimental replicates is often not feasible. This illustrates that the proposed bootstrapping technique offers an accurate way of determining robust connections over a more traditional method, even with limited number of experimental repeats.

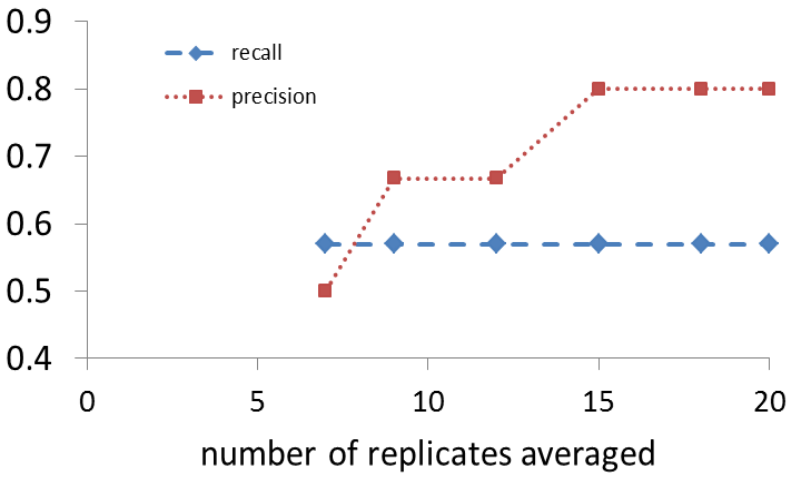

Figure 2.6. Deterministic approach to network identification under noisy data

Increasing number of replicates were generated using 10% noise from the *in silico* results, averaged, and used in the network identification algorithm, with their recall and precision quantified. Although three and five replicates were also used, these are not shown because no solution was found.

2.3.2 Case study 2: ten gene network model

In this example we investigate the performance of the developed algorithm in a larger network consisting of ten genes, as depicted in Equation 2.7.

$$
\begin{aligned}
\dot{X}_1 &= 1 - X_1 \\
\dot{X}_2 &= 1 - X_2 \\
\dot{X}_3 &= X_2^2 - X_3 \\
\dot{X}_4 &= X_3 - X_4 \\
\dot{X}_5 &= X_4^{-1} - X_5 \\
\dot{X}_6 &= 3X_2^2 - 3X_6 \\
\dot{X}_7 &= X_4^{-1} - X_7 \\
\dot{X}_8 &= X_7 - X_8 \\
\dot{X}_9 &= X_{10} - X_9 \\
\dot{X}_{10} &= X_4^{-1} X_5^{-1} - X_{10}
\end{aligned}
\tag{2.7}
$$

For each gene a total of 20 time points of the temporal profile were generated by numerically integrating the coupled ODE system. Integration limits were t={1,21} (a.u.), with initial conditions of 1.9, 1.3, 1.5, 1.5, 0.7, 1.7, 0.4, 1.1, 0.1, 0.55 for genes 1-10, respectively. For the deterministic case study the tolerance was specified at a low value of 10^{-5}. Because the 10-gene network increases the number of binary variables in the upper level to 100, more GA generations are needed to obtain a converged solution; therefore, the number of generations was increased to 1000. The identified connections and kinetic parameters are shown in Figure 2.7(a), with the kinetic orders (g_{ij}) depicted over the connections and kinetic rate constants (α_{ij}, β_{ij}) in brackets over the genes. The comparison of actual and identified time series profiles is shown in Figure 2.7(b). As evident from the figures, the algorithm correctly identified all the connections, kinetic orders and rate constants with a precision and recall of 1.0, thus verifying the satisfactory performance of the algorithm in larger systems.

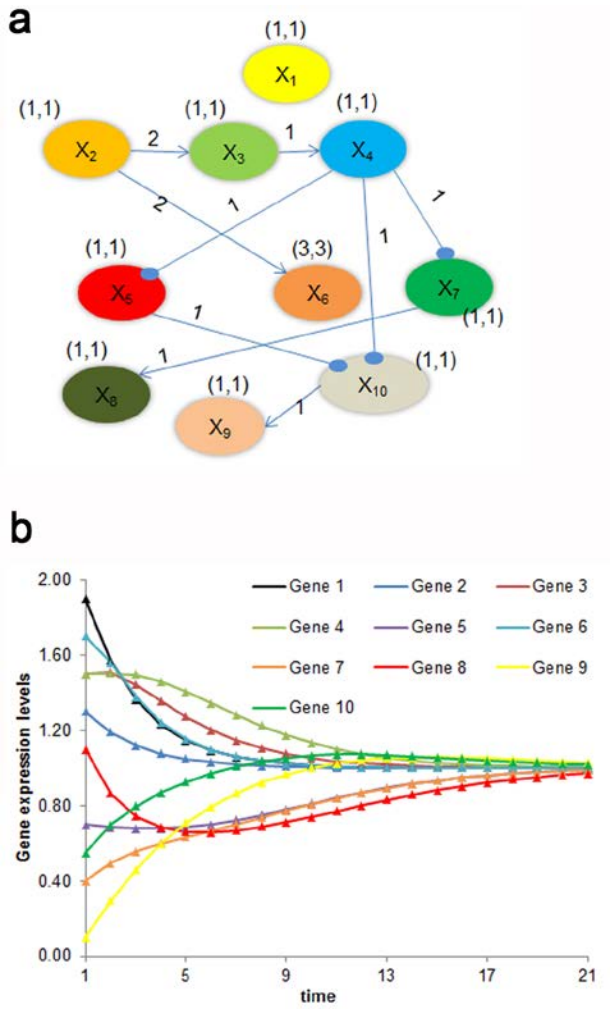

Figure 2.7. Results from the 10-gene network

(a) Identified network. Arrows represent the positive regulation and the filled circles represent the negative regulation of the genes. The kinetic orders of each connection are represented above the corresponding connecting lines and the rate constants for each gene are shown above the genes. (b) Time profile for the ten gene network. The triangles represent the profile generated from the *insilico* data and the lines represent the predicted profiles

2.3.3 Case study 3: experimental data of E. Coli SOS DNA repair

The proposed algorithm is next applied to the SOS DNA repair system of *E.Coli* [81] based on the gene data measured by Ronen *et al.* [82] which is available online [83]. In this model system, the response to DNA damage is governed by a few key genes, which in turn regulate the expression of more than 30 genes which have specific roles in DNA repair. A proposed model is that the RecA protein binds to single stranded DNA, and this nucleoprotein is integral in LexA cleavage, a transcription factor which is a major regulator of the DNA repair genes [81]. The work of Ronen *et al.* investigates the Michaelis-Menten kinetic parameters associated with promoter activity for eight of the major genes in this system. Experimental kinetics were measured by first incorporating a GFP reporter plasmid for each gene's promoter. DNA damage was induced, and the resulting GFP intensities were measured. The number of GFP molecules is proportional to the promoter activity, and can be taken to be analogous to the rate of transcription [82]. We therefore used this promoter activity data [83] to represent gene expression (with the experimental intensity data normalized by the mean column intensity) and used it in our algorithm. Among the four data sets provided by the authors, we chose the third and fourth for this case study because these are measured at the same conditions. For each gene, 29 time points (excluding initial conditions) were utilized. Our objective was to identify regulatory interactions between six genes: *uvrD, lexA, umuD, recA, uvrA* and *polB.*

Identification of this 6 gene network will require 36 binary variables; hence the GA parameters were retained similar to our first case study presented earlier: 20 populations evolved through 200 generations. The error tolerance, however, had to be relaxed to a higher value of 7 because of noise inherent in experimental data set. Figure 2.8(a) compares the actual

experimental data with the predicted profiles generated from the identified algorithm, which shows excellent agreement.

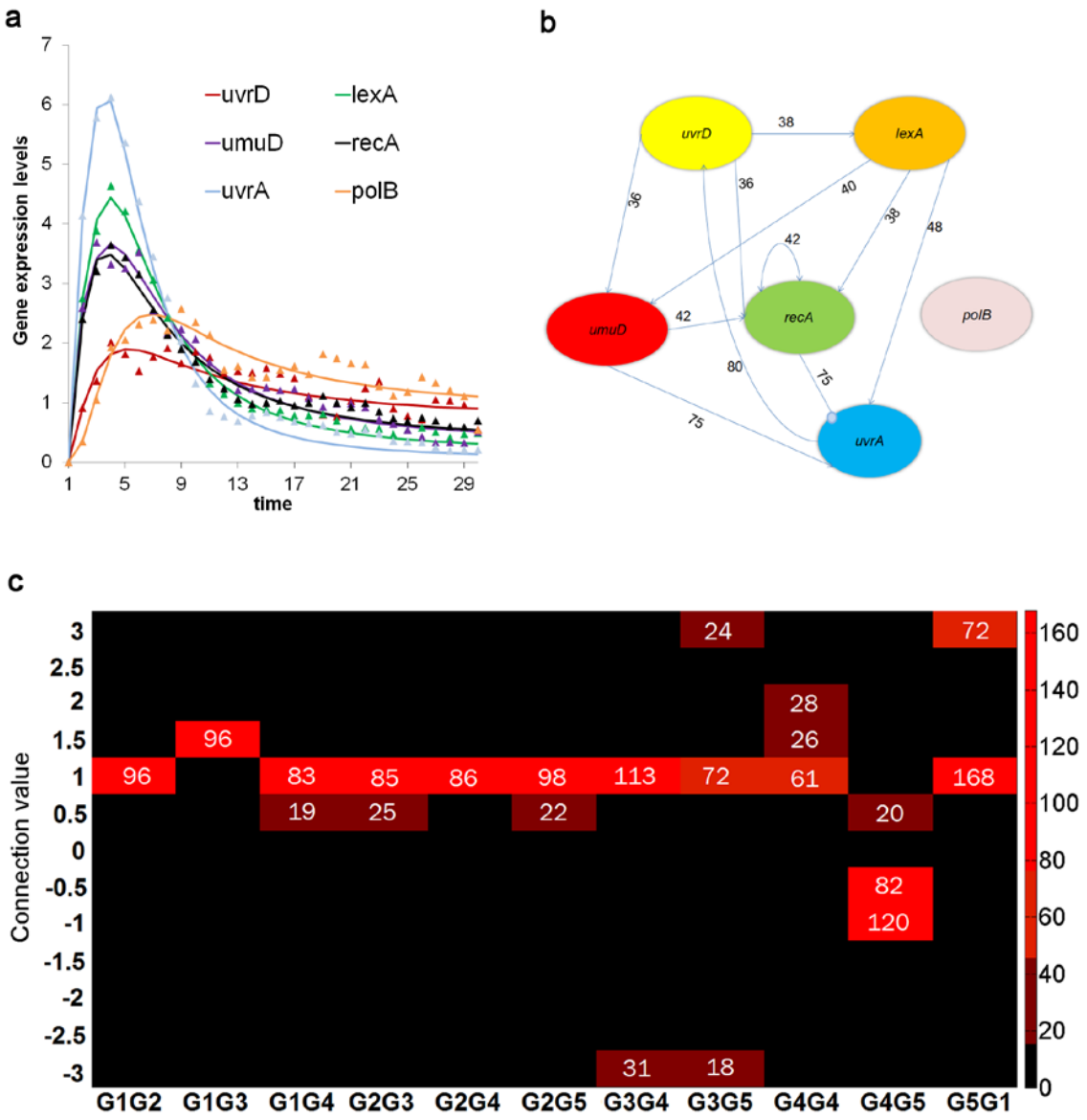

Figure 2.8. Results from the 5-gene experimental E. Coli data

(a)Time profile for the gene network, based off of the mean experimental data. The triangles represent the experimental data and the lines represent the predicted profile. (b) Identified network structure from experimental data for the six gene system. The percentage of connections in the bootstrapping samples are marked on the connections. (c) frequency of specific connection values shown as a heat map (connection coding: 1-uvrD, 2-lexA, 3-umuD, 4- recA, 5-uvrA, 6-polB)

In the next step the robust connections of the identified network are further analyzed by bootstrapping the experimental data set. Since our previous analysis on the first case study demonstrated 200 bootstrap samples to be adequate, in this example we generated 300 artificial data sets from the original experimental repeats. The network identification algorithm was solved for each of the data sets to generate 300 alternate networks. The frequency of occurrence of each network connection is analyzed over the array of alternate network and connections appearing with over 45% frequency are considered to be robust. This bootstrap threshold value was chosen to minimize false negatives and positives (maximize recall and precision), as per the analysis (Figure 2.4(b)) for the similar 5-gene system. Figure 2.8(b) illustrates the predicted robust network for the *E. Coli* data set along with the frequency of repeat of each connection. The corresponding estimated kinetic orders (g_{ij}) and rate constants (α_{ij}, β_{ij}) with 90% confidence level are shown in Table 2.3. The heat map in Figure 2.8(c) further shows how well the algorithm identifies a robust network.

Table 2.3. Results of E. Coli network identification

(a) Reaction orders. (b) Rate constants. (connection coding: 1-uvrD, 2-lexA, 3-umuD, 4-recA, 5-uvrA, 6-polB)

a

Connection	$g_{estimated}$
G1G2	0.9±0.04
G1G3	1.4±0.03
G1G4	0.9±0.04
G2G3	0.9±0.04
G2G4	0.9±0.07
G2G5	0.8±0.08
G3G4	1.0±0.03
G3G5	0.9±0.03
G4G4	1.3±0.07
G4G5	-0.7±0.05
G5G1	1.0±0.14

b Table 2.3 (continued)

Gene	α_i	β_i
X1	3.2±0.28	8.4±0.12
X2	1.5±0.09	1.6±0.19
X3	1.7±0.21	1.5±0.17
X4	5.3±0.16	1.6±0.07
X5	1.6±0.16	2.0±0.15
X6	4.3±0.22	3.8±0.12

2.4 DISCUSSION

The upper level integer programming problem is solved using GA. There are several advantages of using GA to solve the above problem, the most important being that it does not require gradient evaluation. This is a significant advantage for the current network identification algorithm, which contains a nonlinear ODE as a constraint function. In addition, GA starts its search not from a single point in the feasible parameter space, but from multiple locations specified in the starting population. Hence, it holds the chance of converging at a global minimum, although such convergence cannot be guaranteed with GA. However it also suffers from the disadvantage of increased computational cost. All the computations reported here have been carried out on a 2.66 Ghz processer and 16 GB RAM server. The computational time for a five gene network without noise was 1 hour and the same network with noise was 2.5 hours. The computational time for the experimental data was 3 hours. For the 10 gene network, the genetic

algorithm needed more generations to converge, resulting in computational time of 11 hours. Hence extension of the current solution procedure to a much larger data set will be expensive.

In the formulation presented in Equation 2.2, the only user defined parameter is the value of the tolerance which dictates how closely the model prediction must agree with experimental dynamics in order for the network to be considered in the overall algorithm. While for an *in silico* case study without noise the tolerance may not play a vital role, it will be relevant when evaluating noisy scenarios. Specifying a low tolerance value (10^{-3}) in our algorithm under noisy data failed to identify any network, as would be expected. Moreover, using a low tolerance is not advisable when using data sets with noisy replicates since we are not targeting a profile which exactly fits the noisy data; the target is to identify network profiles which describe all the noisy scenarios relatively well. On the other hand a relaxed tolerance runs the risk of compromised prediction quality. In order to quantitatively evaluate the effect of specified tolerance on the identified network structure, the bootstrap/ bi-level optimization algorithm was repeated on the same 5-gene dataset with different tolerance values. Table 2.4 illustrates how the precision and recall of the identified network changes with altered tolerance values. Quite interestingly, it is observed that precision is relatively insensitive to the network tolerance, while recall worsens with increased tolerance. This is very encouraging since this implies that even with relaxed tolerance the identified network does not have false positive connections, although false negative connections increase. Increase in false negatives can be explained by the nature of the objective function, which tries to minimize the number of connections. Hence relaxed tolerance will always lead to a sparser network, as seen in Table 2.4. This analysis indicates that even for a relaxed constraint the algorithm may fail to identify all the connections but the identified connections will always be accurate with low probability of false positivity.

Table 2.4. Effect of error constraint on 5-gene network identification, 5% noise

Error	Precision	Recall	Number of connections
0.13	0.78	1	9
0.2	0.88	0.88	7
0.25	0.78	0.78	7
0.3	0.88	0.7	5
0.35	0.88	0.7	5

In addition to error tolerance, the network identification algorithm performance also depends on the number of data points available for each gene's time series profile (nstep, Equation (2.2)). For the tested five gene network case study, the minimum number of data points needed to accurately capture the network was found to be 20. Interestingly, this number did not change with network size (20 data points per gene were also sufficient in the 10 gene case study), although this number might increase with increased noise.

The performance of the developed robust identification formulation is illustrated using three different systems, both *in silico* and experimental systems representing molecular networks governing gene transcription. Because gene regulatory networks are known to be sparse [67], the present algorithm of minimizing connectivity can be directly utilized. It should be noted that for systems which have been shown to display a high level of redundancy, the current formulation might not be applicable. The first two case studies are based on *in silico* data which allows for detailed analysis of the performance of the algorithm. Overall the algorithm was found to demonstrate excellent predictive capability both in the small 5-gene network along with larger 10-gene network. The proposed bootstrapping scheme was found to adequately capture the precise network from the noisy data as well. Encouraged by the *in silico* results, we applied our algorithm to dynamic experimental data of a 6-gene network responsible for DNA damage repair

in *E. Coli* [82]. While verification of the identified network will be difficult for this system, the time profile of gene expression data predicted by the identified network is in good agreement with the experimental data set. A thorough literature search for existing knowledge of network interactions revealed that quite a few of the predicted connections have been reported in parallel studies. Our algorithm inferred the regulation of *recA*, *umuD* and *uvrA* by *lexA*, which is consistent with the findings reported earlier [84]. Another interesting finding is that our results suggest that *polB* does not influence any of the other genes in the system (*pol B* does not up- or down-regulate any other gene), a finding which was also reported by Kumura *et al.* Furthermore, our identified network shows the self-regulation of *recA*. This protein is the main factor responsible for sensing DNA damage, and has been reported to promote the transcription of itself, thereby promoting damage recognition, and other repair genes [81, 82].

The current approach offers an improvement on existing algorithms. Numerous studies have used the 5-gene network (the current case study I) to test the accuracy and efficiency of their network identification methods. A comparison between the methods is presented by Kimura *et a*l. [84] for the five gene network without noise. While most studies do not report the metrics of precision and recall, the accuracy of the results is still commented on. Most methods have a shorter computational time than the proposed method. However, our algorithm is able to predict a perfect network (recall and precision of 1), while the other algorithms deviate from this. Therefore, there is a trade-off between computational time and accuracy, and selection of the most appropriate method for the system of interest should be chosen judiciously. Nevertheless, this comparison shows that recall and precision are an improvement over many existing algorithms when analyzing the 5-gene network. Additional improvements could be made on the current approach to decrease computation time, such as altering the formulism (e.g. avoiding

direct integration of the system of ODE). Furthermore, as the GA itself is parallelizable, the algorithm could be run in parallel, which could significantly decrease computational time.

2.5 CONCLUSIONS

In this chapter, we present an algorithm to identify robust regulatory networks from time profiles of noisy gene expression data. Our identification algorithm is primarily developed on the hypothesis of sparsity of regulatory network connections. In an earlier work the validity of the hypothesis of sparsity was established using a simplified linear ODE representation of gene expression dynamics in a deterministic system. Herein we further advance the algorithm by incorporating more realistic nonlinear representation using an S-system formulation of gene expression dynamics. The identification algorithm is formulated as a bi-level optimization problem in which the upper level solves an integer programming problem while the lower level is a continuous parameter identification problem. Furthermore, we proposed a framework to incorporate noisy experimental data towards identification of a robust regulatory network. This is done by first generating artificial experimental repeats using the bootstrapping technique, followed by solving the identification formulation for each of the bootstrap data sets. From this library of identified prospective networks we isolate the most repeated network connections which we hypothesize to be robust connections, having low variability to experimental noise. These results show that our bi-level integer optimization algorithm is able to effectively identify the topology and connection strength of gene regulatory networks, even when the gene dynamics are nonlinear and noisy in nature. By using the biological trait of sparsity, the algorithm

optimizes the number of connections in the network while maintaining agreement in gene temporal profiles with the experimental input data. Even with uncertainty and noise in the data, our bootstrapping/identification combination was able to identify a robust network in the *in silico* and *E. Coli* case studies.

3.0 IDENTIFICATION OF SPECIFIC ATTRIBUTES OF EXTRACELLULAR SUBSTRATES INFLUENCING ESC DIFFERENTIATION

3.1 INTRODUCTION

The preceding chapter focused on intracellular behavior, and the identification of regulatory networks from gene expression dynamics. Changes in mRNA levels occur from perturbations to the cell, including extracellular cues. These cues, in part, come from the characteristics of the cells' associated substrate. However, identification of influential substrate characteristics poses difficulties because of their complex nature. In the study outlined in this chapter, we developed an integrated experimental and statistical approach to investigate and identify specific substrate features influencing differentiation of mouse embryonic stem cells (mESC) on a model fibrous substrate, fibrin [85]. Fibrin serves as an ideal substrate platform since its microstructural features can be easily modified by appropriate modification of the fabrication conditions [86]. Such modification also affects the substrate macroscopic property of stiffness, thereby providing a means to investigate the relative importance of macroscopic stiffness and microscopic architecture as cues to which the cells respond.

Twelve different fibrin gels were synthesized by varying fibrinogen and thrombin concentration; this led to a range of substrate stiffness and microstructure. mESC were cultured on each of these gels and analyzed for the extent of differentiation by quantifying pluripotency

and germ layer markers with qRT-PCR. The network topology of the fibrin gels was then characterized by analyzing the electron micrographs of the gels with an image processing algorithm [87]. The complete network topology, which included nine different microstructural attributes, was quantified, and a subset of these features that were most important in describing the conditions and their variability was identified via principal component analysis. The mechanical response of the gels was also measured by both rheology and atomic force microscopy (AFM). The relationship between these explanatory variables (stiffness and microstructural features) and differentiation was then modeled via a second order polynomial. Inclusion of all explanatory variables into the model would require an immense data set, which is often impractical to obtain in the stem cell system. We therefore employed a combinatorial approach to analyze 2-dimensional subsets of the feature space at a time, with the important relationships being determined through significance tests.

3.2 METHODS

3.2.1 Fibrin gel fabrication, mESC differentiation, and gene expression quantification

Twelve different fibrin gels were fabricated by adjusting the fibrinogen concentration (1, 2, 4, and 8 mg/mL fibrinogen) and fibrinogen to thrombin ratios (1.25, 2.5, and 10 mg fibrinogen/U thrombin). Each of these gels was used as a substrate for mESC differentiation for both 2D and 3D culture. For the former, cells were seeded onto the pre-formed fibrin gel, while for the latter, cells were embedded in the gel by suspending them in the fibrinogen before polymerization ensued. For both 2D and 3D cultures, the cells were cultured for a total of 4 days in DMEM supplemented with 10% fetal bovine serum (FBS). At the end of 4 days of differentiation, cells were harvested and gene expression was analyzed for 16 specific germ layer makers with quantitative polymerase chain reaction (qRT-PCR) on a Stratagene MX3005P. Primer sequences used for the analysis, and the genes' associated germ layer, are shown in Table A1. Gene expression data was reported as fold change values calculated herein as $2^{-\Delta\Delta ct}$, where ct values (the cycle count at which the amplified cDNA values are first detected during qPCR) are first normalized to the housekeeping gene β-actin, and then to the control (undifferentiated cells, $\Delta\Delta ct=((ct_{gene,differentiated}-ct_{actin,differentiated})-(ct_{gene,undifferentiated}-ct_{actin,undifferentiated}))$.

3.2.2 Gel stiffness measurements

To assess the mechanical response of the different fibrin gels, stiffness measurements were taken via two different methods: atomic force microscopy (AFM) and rheology. AFM nano-indentation measurements were performed using the MFP-3D Atomic Force Microscope

(Asylum Research). For all measurements a glass borosilicate sphere (diameter 15.9 micron; Thermo Scientific) was attached to the tip of a commercially available silicon-nitride (Si_3N_4) cantilever with a spring constant (k) of ~.1 N/m. A thermal fluctuation method was used for calibrating the cantilever stiffness [88]. The stiffness of each fibrin gel was then investigated by nano-indentation with indentations made at randomly chosen locations considering approximately n = 16 force indentation curves at 3 locations on the fibrin gel [89-91]. The stiffness modulus was determined by applying the Sneddon model to nano-indentation curves from corresponding gels [89, 90, 92, 93]. Curve fitting of the sample indentation depth with the force applied was conducted for a spherical tip model using the following equation relating the force (f) and the sample indentation size (d):

$$f = \frac{4}{3}\frac{E}{1-\upsilon^2}\sqrt{R}\delta^{\frac{3}{2}} \tag{3.1}$$

where E is the Young's modulus, R is the radius of the spherical indenter, and ν is the Poisson's ratio. The sample indentation (δ) is calculated as follows:

$$\delta = \left(z - z_0\right) - d \tag{3.2}$$

Where z_0 is the initial indentation contact point, z is the position of the piezo-electric cantilever, and d is the cantilever deflection. Poisson's ratio was approximated to be 0.5 for all experiments. Curves were fit to small indentation in comparison to the thickness of the samples. The apparent Young's Modulus was obtained by fitting the force-indentation curves to Equations

3.1 and 3.2 with the initial deflection point and Young's modulus (E) as the fitting parameters [94].

To assess the rheological properties, the fibrin was allowed to gel on glass slides and submersed in differentiation media. The slides were then secured to the Peltier cell of a stress-controlled rheometer and the gels subjected to an oscillatory strain, with the stress required to achieve the strain being determined. The storage modulus was then determined from this data.

3.2.3 Fiber network imaging and microstructural characterization

To determine the microscopic structural characteristics of the fibrin, the gels were analyzed by scanning electron microscopy (SEM). First, excess water was removed from the fabricated fibrin gels through serial ethanol dilutions. The ethanol was removed while preserving the fibrin structure through critical point drying (CPD) with CO_2. Samples were then sputter coated with palladium on a 108 Auto Sputter Coater with subsequent imaging on a Philips XL30 field emission gun SEM. Three different images were taken for each fabrication condition, selected at random points on the gel.

To characterize the SEM images, an image based structural analysis algorithm, which has been previously implemented and described [87], was utilized. In brief, a cascade of image processing steps including local thresholding segmentation, morphological processing, and Delaunay triangulation was adopted to identify and associate an artificial struts and nodes network to the real material fibers network. This approach has been qualitatively and quantitatively demonstrated [87] and adopted on a variety of engineered constructs [95-97]. Nine different fibrin gel topological attributes were quantified: pore size, fiber node density (node being determined by the intersection of two or more fibers; density being nodes per unit area),

connectivity (number of fibers per node), pore orientation, fiber orientation, pore aspect ratio, fiber length (length between nodes), fiber diameter, and bulk porosity. The pore and fiber orientation are reported as an index. The orientation index provides a measure of how an angular distribution is concentrated around a specific direction. The fiber index is calculated as:

$$\frac{\sum_{i=1}^{n} \cos^2(\theta_i)}{n} \tag{3.3}$$

where n represents number of fiber segments and θ represents the angle between the segment and assumed alignment direction [87]. This formulation offers a dimensionless number ranging from 0.5 for purely isotropic structures characterized by a random angular distribution to 1 for a set of objects oriented parallel to a specific direction. Therefore, this fiber orientation index (OI) gives a measure of fiber alignment in the system. For pore orientation, an ellipse was associated to each pore and the major axis was considered to calculate a single pore angle with respect to the horizontal direction. For those features which were not scalars (e.g. fiber diameter output being a histogram of the diameters of all fibers in a given image), the mean value was taken. All results reported herein are from analyses using the mean data. This image analysis was performed on the three separate images taken.

Once the fiber network topology was quantified with the image processing algorithm, a principal component analysis (PCA) was performed on the data. This analysis created a new set of orthogonal variables each being a linear combination of the microstructural features, and allowing the elucidation of how system variance was distributed. PCA was performed with the *princomp* function in MATLAB.

3.2.4 Predictive model, regression, and statistical analysis

To describe the relationship between stiffness and the response variable (gene expression), two different approaches were used. First, the strength of the linear relationship was quantified via calculating the Pearson product-moment correlation coefficient. In the second approach, a second order polynomial was utilized, modeled as:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 \tag{3.4}$$

which was able to capture linear as well as nonlinear effects. Equation 3.4 was used to model stiffness (x) versus gene expression levels (y), and was therefore 1-dimensional in the feature space (k=1, k being feature space dimensionality).

The various microstructural features were expected to exhibit cooperative influence towards cell behavior. Incorporating these into one regression model to analyze the cooperative influence of all parameters would introduce a large number of variables which would not be feasible to estimate with the limited availability of data points. We addressed this restriction via a combinatorial approach in which two microstructural features were analyzed per regression. This model was therefore a 2-dimensional (k=2) 2^{nd} order polynomial, shown in Equation 3.5. In the equation, y is the gene expression and x_i is the i^{th} feature. This regression model was analyzed for each combination of features; as there were 9 features which were quantified, a total of 36 combinations were analyzed, as shown in the equation.

$$\begin{bmatrix} y = \beta_{1,0} + \beta_{1,1}x_1 + \beta_{1,2}x_1^2 + \beta_{1,3}x_2 + \beta_{1,4}x_2^2 + \beta_{1,5}x_1x_2 \\ y = \beta_{2,0} + \beta_{2,1}x_1 + \beta_{2,2}x_1^2 + \beta_{2,3}x_3 + \beta_{2,4}x_3^2 + \beta_{2,5}x_1x_3 \\ \vdots \\ y = \beta_{36,0} + \beta_{36,1}x_9 + \beta_{36,2}x_9^2 + \beta_{36,3}x_{10} + \beta_{36,4}x_{10}^2 + \beta_{36,5}x_9x_{10} \end{bmatrix} \quad (3.5)$$

Using the above model, regression was performed to estimate the unknown coefficients (β) by fitting the model to the experimental data (MATLAB).The experimental inputs (stiffness or microstructural feature values) were standardized by centering (by the mean) and scaling (by the standard deviation). Once the regression was performed, statistical tests were carried out to check significance of the overall correlation ($p \leq 0.05$, based on the F distribution). This analysis was done for each gene (16) and for each condition (2D and 3D), giving a total of 64 regressions using Equation 3.4 (16 genes x 2 conditions x 2 stiffness data sets (AFM and rheometry)) and 1152 regressions using Equation 3.5 (16 genes x 2 conditions x 36 feature space combinations). The flow diagram of this screening method is shown in Figure 3.1.

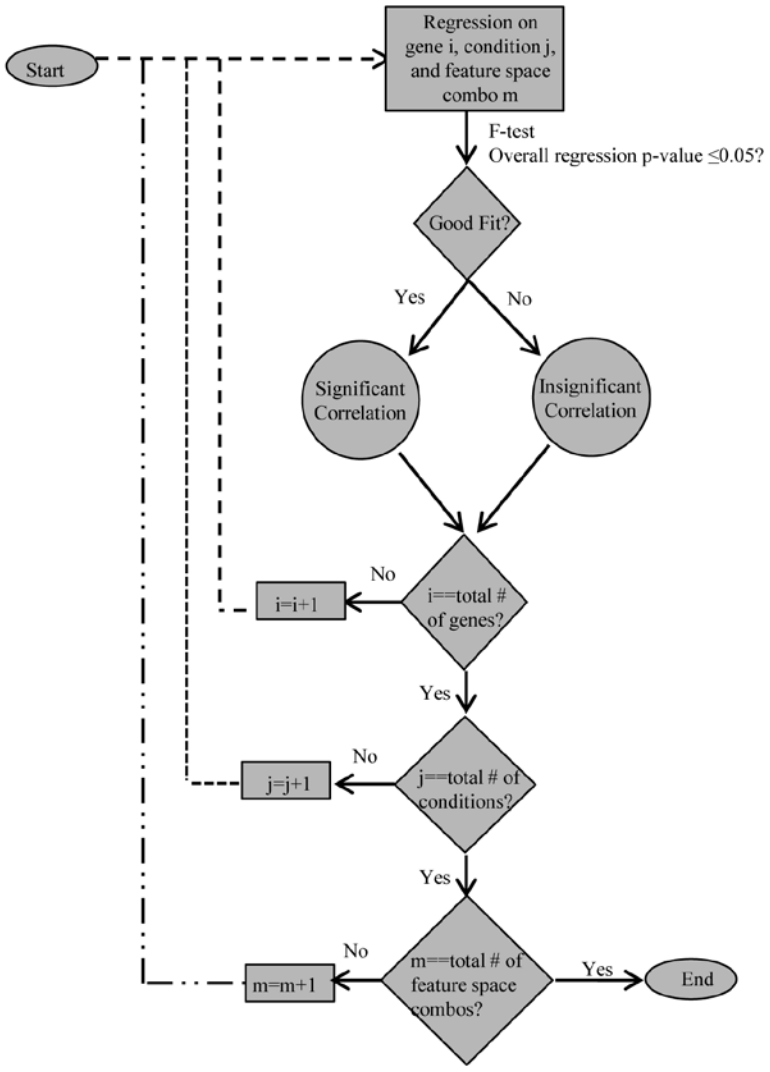

Figure 3.1. Flow diagram of regression and screening methodology to determine significance

3.3 RESULTS

3.3.1 Fibrin gel stiffness and mESC differentiation

Fibrin was fabricated under 12 different conditions by varying the amounts of fibrinogen and thrombin. The stiffness of the fibrin gels at each of the 12 fabrication conditions as determined by AFM and rheometry are shown in Figure 3.2. The overall trend observed was that the stiffness increased with increased fibrinogen and thrombin. mESC were cultured in the synthesized gels both on the gels (2D condition) or embedded in the gels (3D condition).

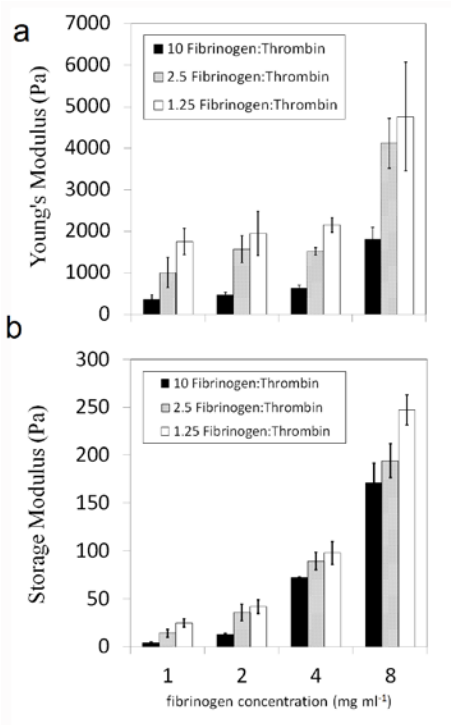

Figure 3.2. Fibrin gel elasticity measurements across various fabrication conditions

(a) Young's Modulus measured by AFM. (b) Storage Modulus measured by rheometry. Legend: mg fibrinogen/U thrombin

After four days of culture, the cells were analyzed for pluripotent and germ layer gene expression. Day 4 was chosen as an appropriate time point to harvest the cells, since the first stage of differentiation normally involves a 4-6 day protocol [98, 99]. At this point the cells are committed to a specific germ layer, and further differentiation to subsequent developmental stages can commence. We have previously thoroughly characterized the behavior of the mESC on these substrates [100]. In the current study we are investigating specific features of the substrate which are influencing differentiation. This requires a thorough, quantitative, and sensitive characterization of the differentiated cells, which we perform by qPCR. Figure 3.3 presents a heat map comparing the expression levels of representative pluripotency and germ layer (endoderm, mesoderm, and ectoderm) markers at different fibrin gel fabrication conditions, for both 2D and 3D cultures. For each row (gene), the expression values are normalized by the gene's maximum level. The greater the dependency of gene expression on fabrication condition, the greater the color variation. While it is observed that most genes are affected to some extent by varying fibrin gel conditions, it seems as if the endoderm gene expression levels change more strongly. In the 2D condition (Figure 3.3(a)) the general trend seems to be higher expression for ectoderm and pluripotency genes with conditions fabricated with more fibrinogen. Interestingly, this trend seems to be reversed in the 3D condition (Figure 3.3(b)). From this figure, it is difficult to discern clear trends for the ectoderm and mesoderm genes. A more analytical approach was therefore taken to determine possible relationships between fabrication conditions and gene expression. In order to quantitatively analyze the effect of stiffness on differentiation patterning, the strength of the linear relationship between AFM/rheometry measurements and gene expression was determined via the Pearson product-moment correlation coefficient. Table 3.1 reports the significance values of the correlation coefficients, which represent how significant

this linear correlation was. Most genes did not show a statistically significant linear correlation between substrate stiffness (either AFM or rheometry measurements) and expression levels. Of the few significant relationships, the majority are for pluripotency genes: *OCT4* and *REX*. Very few of the germ layer markers showed a strong linear correlation with substrate stiffness. In order to examine the presence of nonlinear correlations, 2^{nd} order polynomial regression (Equation 3.4) was implemented, and the significance values of the overall regression were determined (Table 3.1). Still, only a few genes showed significance with the quadratic model (those genes which showed significance in the linear model); this alludes to the possibility that other features of the fibrin substrates, other than stiffness, might have a stronger influence on differentiation.

Table 3.1. Regression significance for elasticity relationship

p-values for correlation coefficients and 2nd order polynomial regressions for the relationship of fibrin gel elasticity and gene expression. Values calculated for both 2D and 3D conditions, with significance ($p \leq 0.05$) denoted by an asterisk (*). Elasticity measured by both AFM and rheology.

	AFM				rheology			
	correlation coefficient		2nd order regression		correlation coefficient		2nd order regression	
gene	**2d**	**3d**	**2d**	**3d**	**2d**	**3d**	**2d**	**3d**
rex1	0.5764	0.0154*	0.5416	0.0441*	0.1789	0.0014*	0.4237	0.0009*
oct4	0.0025*	0.1859	0.013*	0.4369	0.0024*	0.0069*	0.0104*	0.004*
sox2	0.7729	0.2198	0.9585	0.4786	0.1792	0.1476	0.1633	0.1344
brach	0.3321	0.8619	0.6304	0.9858	0.3488	0.6504	0.6336	0.4936
fgf8	0.6048	0.5994	0.8363	0.6928	0.1277	0.0993	0.3332	0.196
gsc	0.3124	0.2509	0.5681	0.5108	0.3411	0.0123*	0.3902	0.0031*
sox17	0.8181	0.1321	0.9675	0.3229	0.9162	0.1	0.8771	0.1656
afp	0.8672	0.7262	0.6726	0.9156	0.3994	0.2465	0.4794	0.2559
hnf4	0.3911	0.6683	0.5988	0.7135	0.1876	0.1441	0.4382	0.2854
fgf5	0.084	0.859	0.1139	0.9826	0.0582	0.3938	0.0732	0.6853
bmp4	0.1864	0.3583	0.1239	0.5809	0.6757	0.0488*	0.8429	0.0241*
cxcr4	0.1007	0.0932	0.2799	0.241	0.0046*	0.0152*	0.0054*	0.0443*
foxa2	0.4988	0.3423	0.5617	0.6138	0.5383	0.1514	0.5406	0.221
nest	0.1452	0.8587	0.3616	0.8613	0.7292	0.9013	0.517	0.8961
ttr	0.343	0.9349	0.4852	0.5058	0.0157*	0.7881	0.0055*	0.9612
gata4	0.2656	0.6815	0.471	0.9	0.3861	0.4862	0.3054	0.4164

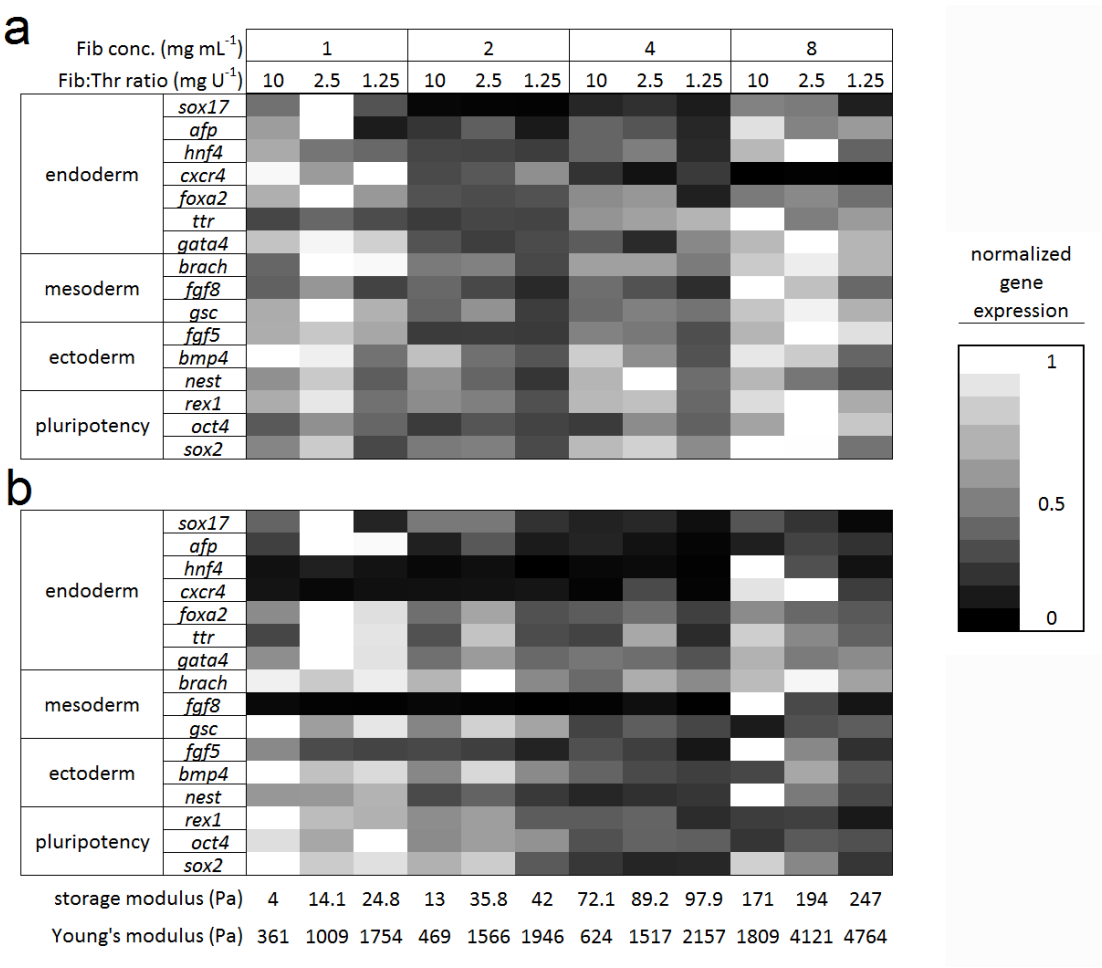

Figure 3.3. Heat map of relative gene expression with fibrin gel

Heat map displaying relative gene expression data of various genes (vertical axis), representing markers for pluripotent cells and each of the 3 germ layers, across different fibrin gel cultures, shown as fabrication condition and gel stiffness measured by rheometry (horizontal axes). Top horizontal axis represents fibrin fabrication condition by its fibrinogen concentration and fibrinogen to thrombin ratio. Data measured by PCR represents cells harvested at day 4 of culture in either 2D (a) or 3D (b) fibrin conditions. For each gene, expression levels were normalized to the gene's maximum level across all gel conditions. Expression values calculated by $2^{-\Delta\Delta ct}$ method, where ct values are first normalized to the housekeeping gene β-actin, and then to the control (undifferentiated cells).

3.3.2 Microstructural features of fibrin gels

Fibrin substrates are highly fibrous in nature, and since the cells are directly interacting with the substrate microstructure, we hypothesized the fibrin microstructural features to be influential in directing cell fate. In order to analyze this further, we first examined the network topology of the synthesized substrates from SEM images. Representative SEM images of gels fabricated at three different conditions are shown in Figure 3.4(a-c).These conditions represent a wide range of stiffness and resulting differentiation behavior (Figure 3.4(d,e)). Even a qualitative comparison revealed differences in the substrate microstructure. In order to accurately determine the differences in the microstructural topology between the fabrication conditions, the fibrous gel characteristics needed to be quantified. The image processing algorithm described in 3.2.3 [87] was applied to the SEM images of all fibrin conditions (three images per gel condition) with the output being quantification of the fibrin topology. Figure 3.4(f-i) shows the algorithm output of four fibrous attributes for the gel conditions shown in Figure 3.4(a-c). There are clearly significant differences in the fibrous features between the three conditions. An interesting observation was that the microstructural features and macroscopic properties are not necessarily linearly related. For example, the gels fabricated under the first two conditions in Figure 3.4 gave similar stiffness (d). However, microscopic comparison of these two gels shows differences in pore size and fiber length (a, b, f, h). Gene expression resulting from culture on these two gels is also different, supporting the hypothesis that factors other than stiffness might be more influential in guiding differentiation. Examination of gels at a higher stiffness (Figure 3.4(c)), also showed microstructural differences from the softer gels, but these differences vary depending on the type of feature and fabrication condition, further demonstrating the complexity of the system.

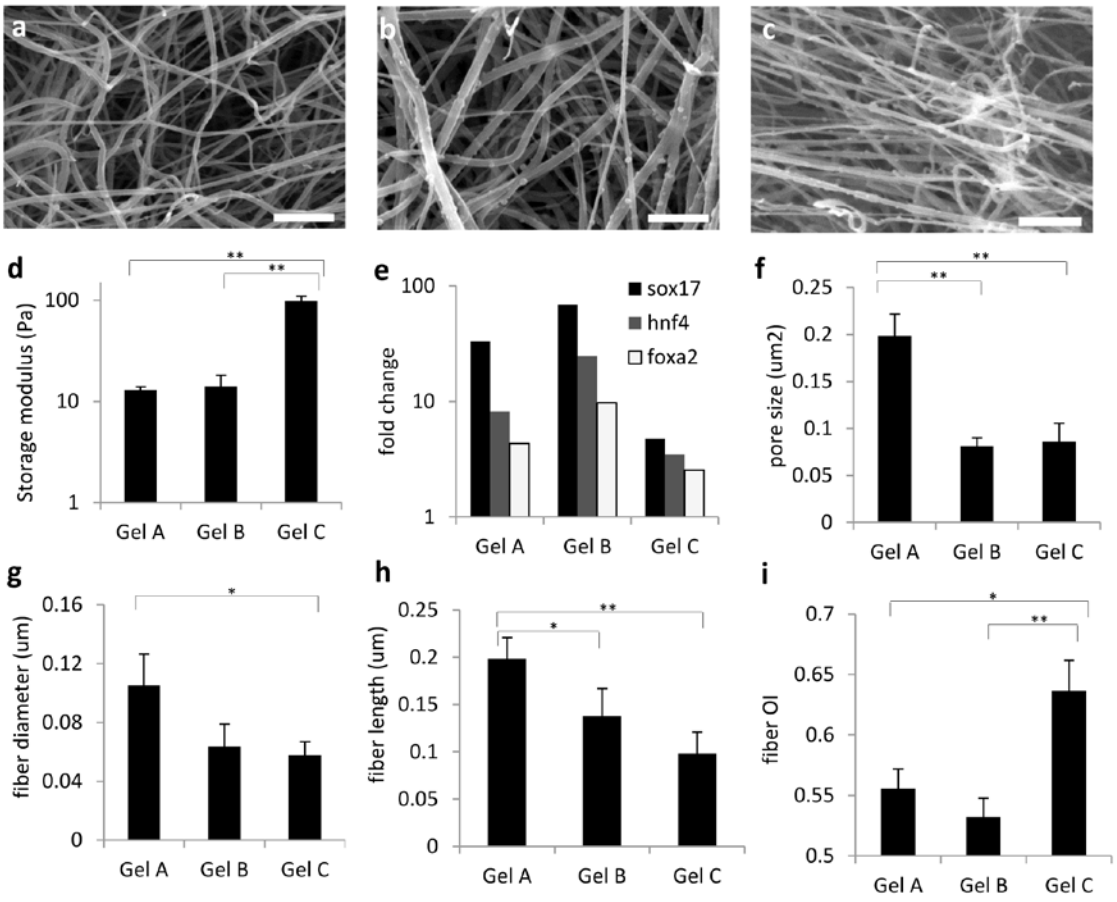

Figure 3.4. Variable characteristics and behavior associated with fibrin gels fabricated under different conditions

(a-c) SEM images of fibrin gels fabricated with 2 mg/mL fibrinogen and a fibrinogen to thrombin ratio of 10 mg/U (a), with 1 mg/mL fibrinogen and a ratio of 2.5 mg/U (b), and with 4 mg/mL fibrinogen and a ratio of 1.25 mg/U (c). Bar = 1 μm. (d) Stiffness of these gels as measured by rheometry. (e) Differentiation patterning of cells cultured for 4 days in the aforementioned fibrin gels, 3D condition. Fold changed calculated by the ΔΔ Ct method as described in section 3.2. (f-i)Various fibrin fibrous attributes for the three gels quantified by the image processing algorithm. In (d-i), x-axis labels represent fibrin gels fabricated under different conditions: 'Gel A' with 2 mg/mL fibrinogen and 10 mg/U fibrinogen:thrombin ratio; 'Gel B' with 1 mg/mL and 2.5 mg/U; and 'Gel C' with 4 mg/mL and 1.25 mg/U. The average values of the features are presented. In all graphs significance performed via a Student's t-test. * $p<0.05$, ** $p<0.01$.

To quantify the entire microstructural topology, the image processing algorithm was applied to the SEM images of all 12 gel conditions. Figure 3.5(a) displays the fibrous network, identified by the algorithm, for a representative image. The algorithm does an excellent job identifying the individual fibers and their connections. Two representative topological attribute histograms generated from this identification are shown in Figure 3.5(b) and (c) for fiber diameter and pore area respectively. This image and quantification reveals differences in the distribution of these two fiber network features: fiber diameter was approximately normally distributed; pore distribution, while dominant in a narrow range of small pores, still contained numerous outliers of large area. Nine different attributes were likewise quantified for 3 different images of all the 12 gel conditions to characterize the network, as outlined in section 3.2. How these attributes change with varying fibrin fabrication conditions is represented in Figure 3.5(d-l), where the data represents average over 3 representative images.

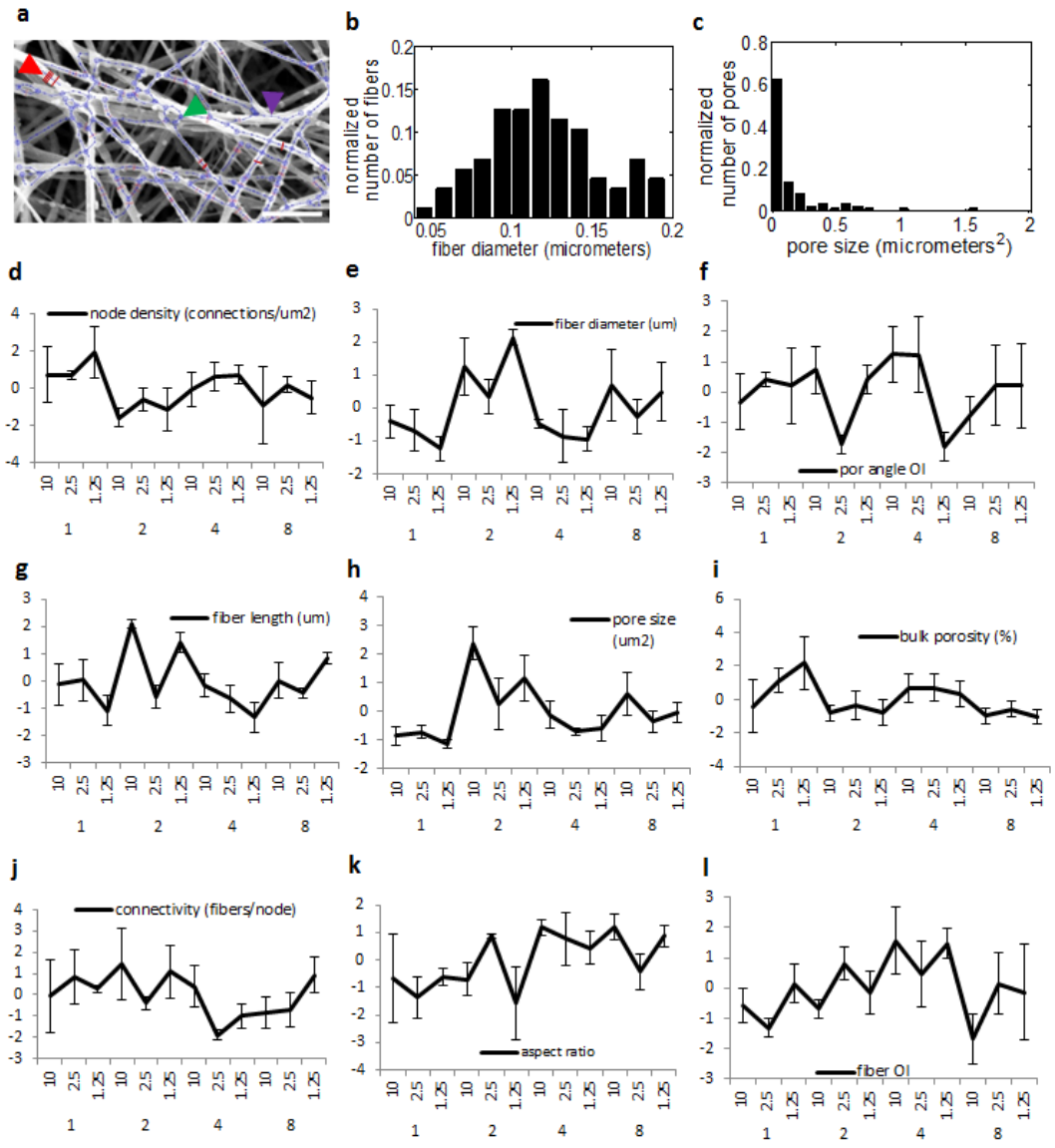

Figure 3.5. Results of the image processing algorithm applied to the SEM images of the different fibrin gels

(a) SEM image and identified fiber network. Lines traversing fiber width (red arrow head) indicate detected diameters, with several of the lines in bold for clarity; lines running along the fiber length (purple arrow head) denote identified fibers; large circles (green arrow head) indicate nodes. Bar = 1 μm; (b,c) output histograms of fiber diameter and pore size, respectively (a-c are representative outputs for fibrin gels fabricated with 2 mg/mL fibrinogen and 1.25 mg fibrinogen/U thrombin); (d-l) Compiled output (standardized, by centering by the mean and scaling by the standard deviation) for all 9 features (error bars represent 1 standard deviation, n=3 images). Categories on horizontal axes represent fibrin fabrication condition: top row-mg fibrinogen/U thrombin; bottom row-mg fibrinogen/mL. Fiber length represents distance between two nodes. OI denotes orientation index.

These attributes were further analyzed by principal component analysis (PCA) in order to quantitatively determine the microstructural features more sensitive to the gel fabrication conditions. This data reduction technique, performed via singular value decomposition of the data, leads to generation of orthogonal variables and subsequent identification of variables that capture maximum data variance. .The resulting biplot (Figure 3.6) shows the contributions of each of these features towards the first and second principal components. By projecting the feature vector values onto the x axis, one can determine the importance of the features to the first component, which is responsible for most of the system variability. The most influential features are the ones that show the largest projection magnitude. Considering 0.1 as a lower threshold, features which contribute the least to feature space variability were identified to be pore angle orientation index and pore aspect ratio, which were excluded from subsequent analyses. The more influential features which were retained for subsequent analysis were fiber length, fiber diameter, pore size, node density, porosity, connectivity, and fiber orientation index (fiber alignment).

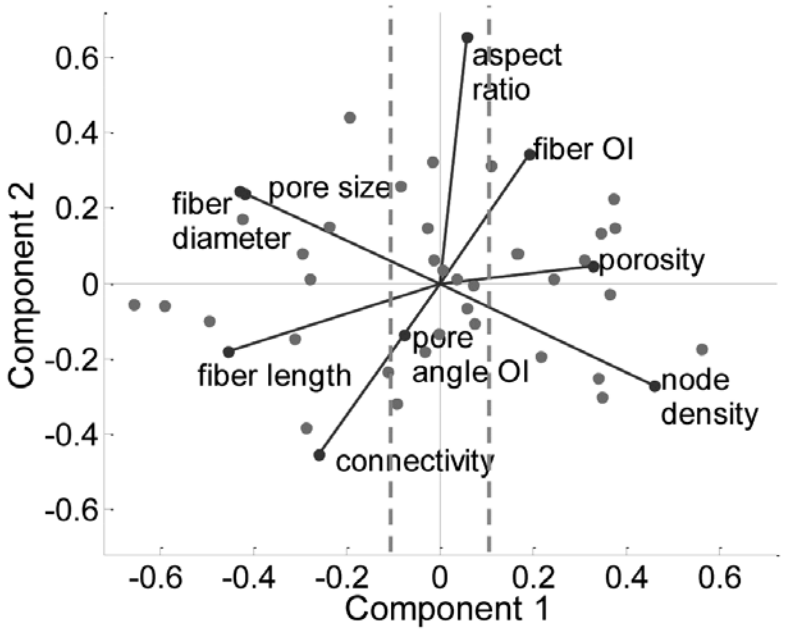

Figure 3.6. Identified influential microstructural features

Biplot of the principal component analysis (PCA) of the feature space. Filled circles represent observations in the principal component space, while vectors indicate the contributions of each feature to the 1st (x-axis) and 2nd (y-axis) principal component. OI denotes orientation index. Vertical dashed lines represent the 0.1 magnitude threshold value of the 1st principal component coefficients used to screen important features.

3.3.3 Correlating microstructural features and differentiation

Having identified the dominant microstructural features sensitive to the gel fabrication conditions, the next task was to analyze the gene expression data with respect to the gel microstructure, and to determine if specific genes were strongly correlated with this feature space. Incorporating all dominant microstructural features into the correlation would require an immense data set. To utilize a limited data set, this correlation was determined by performing a regression analysis for each of the 16 genes against a 2-dimensional second order polynomial

(Equation 3.5), considering combinations of two dominant features at a time (independent variables, x_1 and x_2). The analysis was repeated for both the 2D and 3D culture conditions (see Figure 3.1 for algorithm).The output of each regression was a set of best fit polynomial coefficients and the corresponding regression surface. Shown in Figure 3.7 is an example of two significant correlations relating *FOXA2* (a) and *SOX17* (b) gene expression to fiber length and diameter. The feature space is in its centered and scaled form, transformed as such for the regression analysis. As shown in both surfaces, a minimum seems to occur, as dictated by the 2nd order polynomial. Highest expression is achieved when fiber length is high and diameter low, or vice versa. The power of this analysis lies in the possibility of utilizing the microstructural information to aid in the design of materials that guide stem cells towards desired phenotype fates. By isolating important features of the fibrin substrate and determining how they affect cellular behavior during mechanical differentiation induction, one can try to mimic these features and recapitulate this substrate topology on prospective future synthetic induction substrates, rather than relying on a "guess-and-check" method.

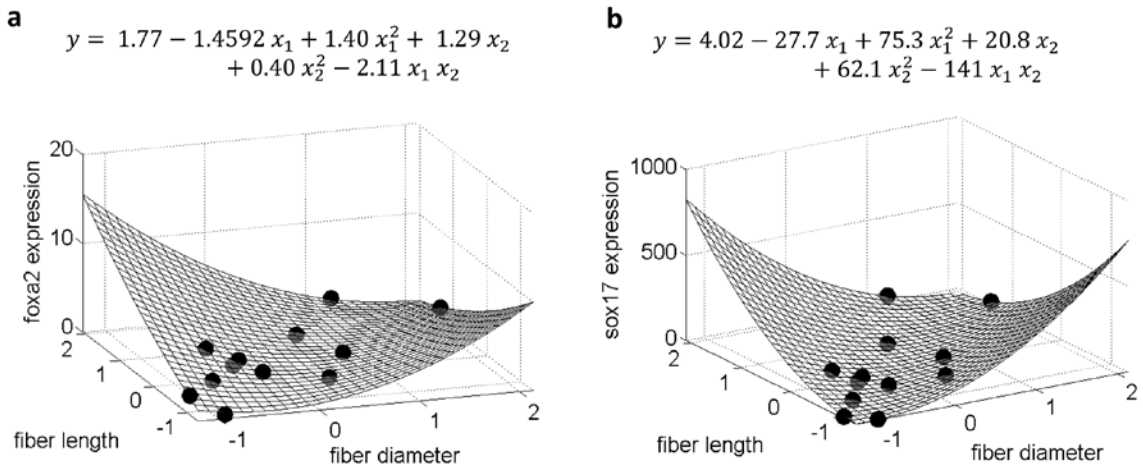

Figure 3.7. Predicted response of representative genes to two features

Regression surfaces and associated best-fit regression polynomial showing the regressed relationship between the influential features of fiber diameter/length and *FOXA2* expression, 2D condition (a) and *SOX17* expression, 3D condition (b). In the equations, x_1 denotes fiber diameter and x_2 fiber length. Expression values are in relative fold change, and features have been centered and scaled (previously described). Circles denote experimental data.

A screening procedure was next implemented to isolate the most significant correlations and parameters. As stated in section 3.2, 1152 regressions resulted from the 9 different feature space, 2 gel condition, and 16 gene combinations. However, in order to isolate the effect of important microstructural features, only combinations of the 7 influential features identified by PCA were considered for the feature space. Therefore, for each gene, 21 regressions per condition (2D and 3D) were analyzed. Out of these regressions, significance was analyzed by determining the p-value of the overall regression. The resulting p-value for 42 regressions (21 x 2 conditions) for each gene is shown in Figure 3.8(a). In the figure, each point represents each individual regression and the resulting significance levels (y-axis). Those correlations which have an overall p-value≤0.05 were considered significant. These significant regressions are

recorded in Table 3.2 and Table A2, which records the microstructural feature combinations which showed strong relationships to specific genes. Figure 3.8(b) displays the number of correlations per gene with p-value≤0.05. As shown, gene expression levels of *REX1*, *BRACHYURY*, and *BMP4* do not show any strong relationships with microstructural features. The genes of *AFP*, *FOXA2*, and *GATA4* show the strongest relationship with microstructural features, having the highest number of significant correlations with the features space (12, 11, and 8 significant correlations, respectively).

Table 3.2. Regression significance for microcharacteristic relationship

p-values for the significant ($p \leq 0.05$) 2^{nd} order polynomial regressions relating two microstructural features to gene expression, 2D condition

gene	feature 1	feature 2	regression p-value
sox2	porosity	connectivity	0.0339
fgf8	porosity	connectivity	0.0208
gsc	fiber diameter	fiber OI	0.0478
nestin	porosity	fiber OI	0.0449
sox17	pore size	fiber OI	0.0117
sox17	node density	fiber OI	0.0044
sox17	fiber diameter	fiber OI	0.0059
afp	pore size	fiber OI	0.0303
afp	node density	fiber OI	0.0028
afp	fiber diameter	fiber OI	0.0039
afp	fiber length	fiber OI	0.0022
foxa2	pore size	fiber length	0.012
foxa2	pore size	fiber OI	0.0425
foxa2	node density	fiber length	0.0476
foxa2	node density	fiber OI	0.0272
foxa2	fiber diameter	fiber length	0.0078
foxa2	fiber diameter	fiber OI	0.0348
foxa2	porosity	fiber length	0.0346
ttr	pore size	fiber OI	0.0041
ttr	node density	fiber OI	0.0216

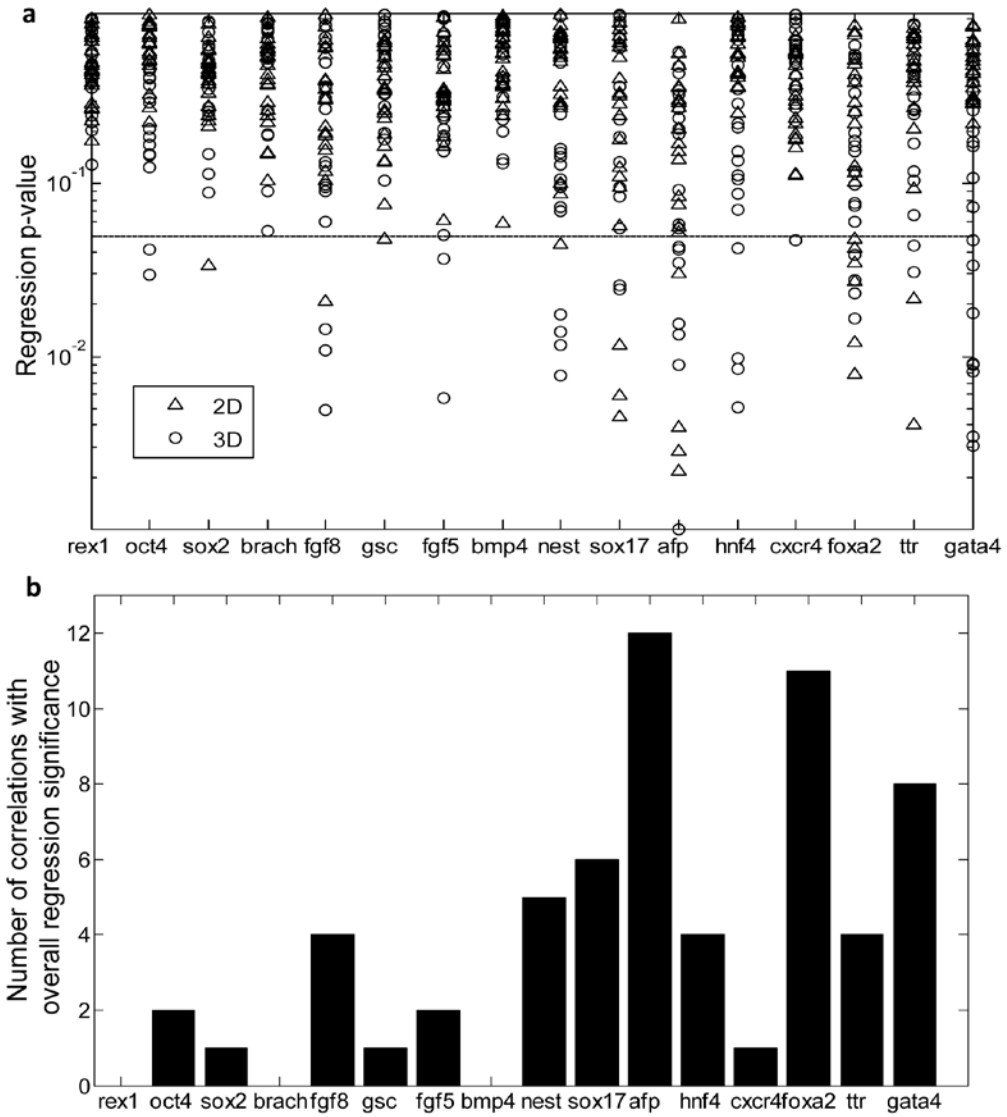

Figure 3.8. Significance levels for each regression

(a) The expression level data for each gene was regressed onto each 2 dimensional feature space combination (using the 7 most prominent microstructural features previously identified), and the overall regression significance level is recorded. Each data point represents a single regression (for each gene, 21 regressions per condition (2D/3D) were performed). Those correlations which had an overall p-value≤0.05 (horizontal line) were considered significant. (b) For each gene, the total number of correlations (both 2D and 3D) having an overall regression p-value≤0.05 were compiled per gene.

3.3.4 Germ layer specificity in response to microstructural features

Next we systematically investigated whether any specific germ layers showed a stronger response to the microstructural feature space. From the screened significant correlations (Table 3.2 and Table A2), the gene markers were segregated, and the results compiled, based on germ layer (see Table A1). Figure 3.9(a)shows the compiled results of the combined 2D and 3D conditions; overwhelmingly, genes from the endoderm lineage showed more significant correlations with microstructural features than those of mesoderm, ectoderm, and pluripotency. The figure shows that the average number of significant regressions per gene for endoderm is at least 2.8-fold higher than the other markers. These data suggest that while the examined microstructural features govern differentiation to some extent for several phenotypes, this effect is much stronger for endodermal genes.

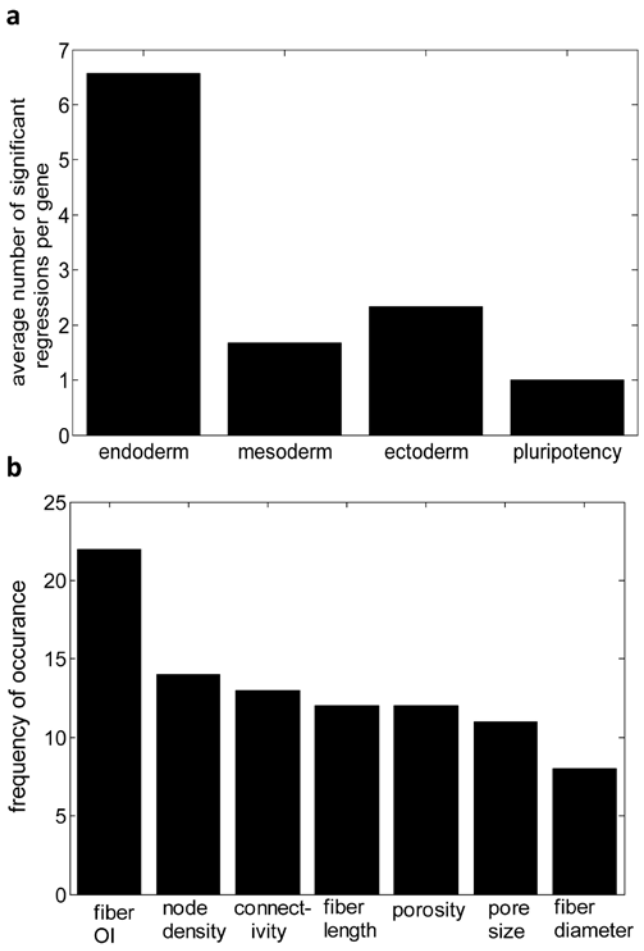

Figure 3.9. Comparison of the effect of microstructural features on gene expression between germ layer/pluripotency markers

(a) The average number of significant correlations per gene is organized for each phenotype. (b) Frequency of occurrence of influential microstructural features in significant endoderm correlations. Results out of a total of 46 significant regressions. OI denotes orientation index (representing fiber alignment).

Endoderm, being identified as the lineage which was most significantly affected by fibrin gel microstructure, was further analyzed to determine if any specific microstructural feature(s) were most prominent in guiding differentiation behavior. The regressions against endodermal gene expression which were found to be significant were first tallied. Of these significant correlations, the microstructural features present in the regressed equations were analyzed for frequency of occurrence. This can be done by analyzing Table 3.2 and Table A2 for only endoderm genes, and determining the frequency of occurrence of each microstructural feature. These results are compiled in Figure 3.9(b). All of the features identified by PCA occurred to some extent in guiding endoderm differentiation, as they were all present to some degree in the significant correlations. However, Figure 3.9(b) shows that fiber alignment (represented by the fiber orientation index) was the most important, as it appeared most frequently, greater than 50% more often than any of the other features.

3.4 DISCUSSION

The most common method to drive embryonic stem cell differentiation is through chemical cues [17, 101, 102]. Recently, however, substrate properties have been reported to modulate cellular behavior, including elasticity and fibrous characteristics [20, 103]. However, identification of specific physical cues in a complex substrate which are most influential in guiding cellular behavior still remains a challenge. In this chapter, we have developed a systems level approach to guide this identification and applied it to analyze the effect of fibrin substrate microstructure on ESC differentiation without the use of any chemical cues. The highly fibrous nature of fibrin and the ease with which its fibrous topology can be adjusted makes it amenable

to investigation of cell-substrate behavior [86, 104]. The fibrous microstructure of fibrin consists of various attributes with which cells are likely to interact, including fiber diameter and orientation [103, 105]. Analysis of such complex interactions and identification of dominant attributes governing cellular behavior requires an integrated experimental and mathematical approach. This was accomplished with a systems level approach incorporating experimental, statistical, and image processing techniques in the analysis. Differentiation was quantified by PCR, and the relationship between differentiation and stiffness/microstructural features determined through a combinatorial statistical modeling approach. This combinatorial approach enabled the utilization of the complete microstructural feature space in the correlation analysis even with limited experimental data. Through this screening method, we were able to elucidate the relative strength of the correlations.

3.4.1 Importance of fibrin and microstructural regression analysis

Fibrin is becoming increasingly popular as cellular scaffolds, and is known to support cell survival, growth and differentiation [106, 107], with its biodegradable nature making it appealing for possible transplantation applications [86, 108]. Fibrin has been studied as a scaffold for mESC-derived progenitor cells [109, 110], and shows promise for future regenerative therapies, as has been shown by fibrin/mESC *in vivo* studies using rats [111] and engineered therapeutic protein delivery vehicles [112]. With more techniques being reported on the modification and control of fibrin substrate properties [113, 114], it may be possible to tailor fibrin substrates to guide cellular behavior, thereby improving its therapeutic uses. This potential could be realized to a larger extent if more information was available on how the substrate affects cellular behavior. Through the current approach, we offer a platform to advance this understanding.

In this study we have shown that during mESC differentiation on fibrin substrates, the microstructural characteristics of the substrate show stronger correlations with differentiation patterning of mESC than stiffness alone. We first tested the relationship of the latter by changing fibrinogen and thrombin amounts to create fibrin gels of varying stiffness, each of which being used to induce differentiation. Increasing fibrinogen and thrombin increased the stiffness of the fibrin substrate, an effect which corresponds to previous reports [104]. Although a trend was qualitatively observed between fibrin stiffness and differentiation (Figure 3.3), a statistically significant correlation was not present between these two variables ($p>0.05$) for the majority of the genes. In addition to stiffness, microstructural features also differed between substrate fabrication conditions (representative images shown in Figure 3.4); we therefore attempted to quantify the fibrin network topology and investigate whether certain microstructural attributes are important in phenotype commitment. While limited research has focused on analyzing the relationship between network topology and cellular behavior, it is particularly important when considering the scale of the interactions: the fiber microstructural features identified herein are on the micron-scale. These features are on the same scale as cellular components, and therefore can facilitate scaffold-cell interactions. Prior studies have examined substrate microstructure and how micro-characteristics affect behavior of cells [103, 105, 115]; however, these studies have typically focused on a few isolated micro-characteristics. We chose to perform a more exhaustive analysis by screening for 9 different topological features, and further expand on the research area by interrogating the effect of these features on stem cell phenotype commitment. It should be noted that the current method is applicable to fibrous network topology, and would not be amenable to more amorphous substrates, although other characterization approaches ([116, 117]) might be useful when identifying fiber orientation and angle distribution in amorphous materials.

During analysis of the fibrin gels, three random points on the fabricated gels were selected for characterization, both during microstructural topology and AFM stiffness quantification. These features were then averaged for the regression studies. A possible limitation might arise if there is significant heterogeneity, i.e. different gel stiffness and fiber topology, within a given gel, which could result in the cause-effect relationship between the fibrin and the cells to change depending on gel location. This heterogeneity, or lack thereof, can be quantified through AFM. This is a convenient tool to analyze heterogeneity and differences in microenvironment due to its precise nature (AFM tip was 15.9 micron) and ability to scan an area. For each gel location analyzed with AFM, up to 16 different measurements were taken. Figure 3.10(a) shows the coefficient of variation (CV) of the AFM measurements for each gel condition. As shown, over half of the gels exhibit stiffness CVs less than 20%, which suggests that heterogeneity is low across the gels. This is further demonstrated through a histogram of AFM values for a representative gel (Figure 3.10(b)). Therefore, the mean values of the fibrin characteristics should be appropriate in the regression analysis. If a more exhaustive scan of fibrin heterogeneity would be required, AFM could be used, with more sample points across more of the gel.

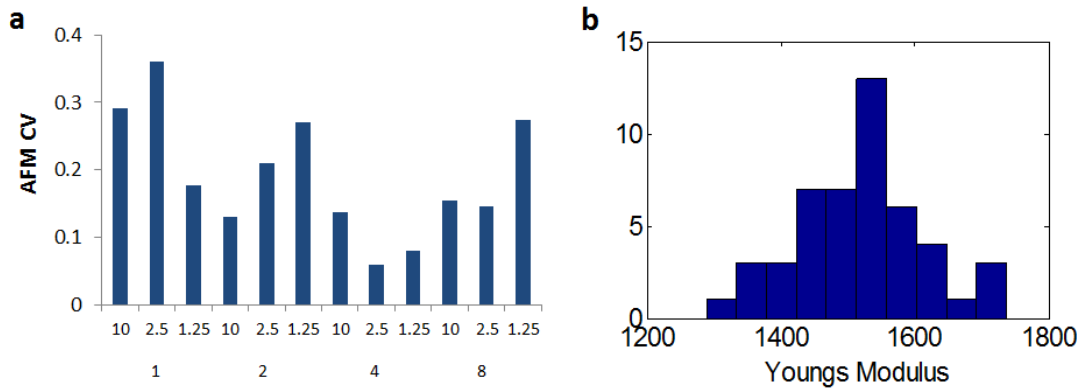

Figure 3.10. Stiffness homogeneity of fibrin gels

(a) Coefficient of variation (CV) for AFM measurements for each gel condition. Categories on horizontal axes represent fibrin fabrication condition: top row-mg fibrinogen/U thrombin; bottom row-mg fibrinogen/mL. (b) histogram of Young's Modulus measurements across different locations of the fibrin gel fabricated with 4 mg fibrinogen/mL and 2.5 mg fibrinogen/U thrombin.

Screening for significant correlations between microstructural features and gene expression was performed via a 2nd order polynomial regression. Developing correlations and mathematical models between these variables, although more complex, was advantageous over simple comparison for numerous reasons. While comparison can give statistical differences between conditions, mathematical correlations can give a better understanding of how the observed variable changes with the features space. Furthermore, optimization can be performed on the model to yield possible optimal responses. Ideally, this polynomial would include all features, with reduction of terms and subsequent selection of influential features being performed by such techniques as ridge regression, lasso, or backward stepwise selection [118]. However, because of the limited amount of data, combinations of two features were selected at a time. It is important to note that a proper design of experiments (e.g. central composite design) can enable more information to be extracted from the system via response surface methodology [119] (for

an example of the use of this technique for investigating cell-substrate interactions, see [120]). This would entail performing experiments at precise points in the microstructure feature space. Because of the nature of our system, the microstructural features could not be precisely controlled, and therefore a design of experiments analysis was not feasible. Furthermore, exact information on the individual parameters is difficult to extract because of possible collinearity of the features (Figure 3.6). However, these restrictions do not prevent the significance level of the overall correlation (using the 2^{nd} order polynomial model) from being obtained and from being used to determine if the microstructural features are strongly correlated to gene expression. In addition, through our screening approach one can obtain a sense of which features are most important, even in the presence of possible collinearity, and can perform further experiments to more accurately determine the effect of those important features on differentiation. For example, Figure 3.9 shows that fiber alignment occurs most frequently in the significant endoderm regressions. Given this, we can speculate that fiber alignment is influential in guiding endoderm differentiation, and further experiments can be done in which fiber alignment is controlled with other features invariant (with such tools as electrospinning and design of experiments). This may give more accurate response surfaces relating fiber alignment to differentiation, thereby yielding information which can be used in substrate design to optimize endoderm differentiation.

3.4.2 Applicability of method to other systems

The response between substrate characteristics and differentiation was taken to be a 2^{nd} order polynomial. If a different form, for instance sigmoidal, existed between the variables, it is possible that the current model would not identify the correlation as significant, even if a strong relationship was present. However, specific functional forms of a response are usually only

considered when theory dictates that specific behavior. When little theoretical information is known, as is the case with the current system, polynomial models are most often used, and are very appropriate to capture first-order, second-order, and interaction behavior. It should also be noted that for those microstructural features giving a range of values per image as opposed to a single value (e.g. fiber diameter vs. bulk porosity), the mean of this range was used in the polynomial model. If the whole of the data were to be used (utilizing the whole histogram as explanatory variable values in the correlation), more sophisticated methods would have to be employed, which is not in the scope of the current work.

In the presented system of mESC on fibrin gels, microstructural features exhibited a strong correlation with differentiation, while that with stiffness was relatively weak. This is in contrast to studies which have shown that substrate stiffness influences differentiation [20, 120-126]. Most of these studies, however, employ synthetic polymer substrates that do not exhibit a discernable microstructure. Indeed, we have previously shown that in the system of mESC on alginate gel, differentiation does seem to be a function of the alginate stiffness [127]. However, this polymer is much more amorphous in nature, and does not have a fibrous microstructure. Furthermore, alginate is inert, and does not directly interact with the cells. In contrast to these studies, the present study has used a fibrous fibrin substrate with which the cells directly interact, and hence are likely to be more sensitive to the microstructure.

3.4.3 Comparison of specific genes and 2D/3D conditions

The relationship between gene expression patterning and microstructural features was determined to be lineage specific. The endoderm germ layer most strongly correlated with fibrin microstructural topology, with the pluripotency, mesoderm and ectoderm germ layer phenotypes

being less responsive to substrate modulation. More specifically, the endoderm genes of *AFP* and *FOXA2* exhibited the most responsiveness to the fibrin microstructure, showing significant responses to the features in both 2D and 3D culture conditions. It is interesting to observe that while the downstream endoderm genes of *GATA4* and *HNF4* showed significant correlation with the fibrin topology, this is only present in the 3D case and not 2D. Further comparison of the different conditions shows that more significant correlations are present in the 3D condition than 2D. The fibrin gel was expected to have less interaction with the cells under 2D conditions, as compared to 3D where the cells are completely embedded in the substrate. Therefore, the topological features of the fibrin could influence cell behavior to a lesser extent on the 2D gels, explaining the fewer significant correlations. While the current analysis shows that these specific genes were most strongly affected by network topology, a more rigorous analysis is needed to show co-regulation. This, for example, could be achieved through biclustering, which we have previously utilized to determine co-regulation across fibrin conditions [128], but could potentially be applied to microstructural features.

3.4.4 Contribution of other substrate factors and mechanistic information

The current work focused on comparing the effect of different fibrin substrate cues on mESC differentiation. Because fibrin is biodegradable, temporal changes in the substrate, in addition to degradation products, could influence differentiation. However, the current study was performed for a relatively short time period of four days, and while some fibrin gel degradation was observed, it was not significant and only becomes somewhat visible towards the very end of the differentiation protocol. Furthermore, this degradation was similar in all of the synthesis

conditions, and would therefore not affect the correlation results significantly. In addition to substrate characteristics, the cellular niche does comprise of other factors, including extra-cellular matrix [129] and varying amounts of fibrin-bound thrombin [130, 131]. While these other factors might influence cellular behavior to some extent, because of the strong correlation found between microstructural features and differentiation, we feel that, in the current system, this relationship is the most influential. There are also media-related factors, including diffusion of ligands and growth factors to the cells, which would affect differentiation. However, since the media is the same between conditions, differences in these factors would be a function of microstructure. For instance, the diffusion of soluble factors to the cells would be a function of fibrin porosity. Therefore, the microstructural features are the independent variables in this system, although their effect on the cells could be considered through modulation of associated factors. The described screening approach and utilization of the image processing algorithm did not require selection of these microstructural features to analyze, but allowed the testing of correlations involving the complete fibrous network topology. Specific features identified to be influencing differentiation are fiber alignment, node density, fiber length, pore size, fiber diameter, porosity, and connectivity, of which fiber alignment was by far the most influential. Interestingly, other investigators have also shown that certain of these features affect cellular behavior, albeit in different systems and in different ways. Fiber alignment has been shown to be important for Schwann cell migration and neurite outgrowth [132]. The proliferation and morphology of osteoprogenitor cells seems to be affected by fiber diameter, while differentiation is not [105]. Fiber diameter and orientation have been shown to affect fibroblast morphology but not proliferation [103]. Herbert *et al*. studied the system of dorsal root ganglia on fibrin gels, and postulated that neurite behavior was governed more by the fibrin density rather than the number

of fibrin bundles or bundle diameter [115]. While these studies showed that fibrous micro-characteristics are important in guiding cellular behavior, the current report is the first systematic study to analyze the effect of fibrin fiber network topology on mESC differentiation.

The proposed systems level analysis offers a rigorous platform to identify and quantify cause and affect relationships. However, it does not provide any mechanistic information of the relationship. Such mechanistic studies have been reported in other systems. Mukhatyar *et al*. investigated the effect of fiber alignment on Schwann cell migration and neurite outgrowth, and determined that more aligned fibers promote fibronectin adsorption which in turn influences these two cellular behaviors [132]. Dalby *et al*. demonstrated the importance of nano-topology on mesenchymal stem cell differentiation: substrate nanoscale disorder promoted bone mineral production, with this behavior postulated to be governed by adhesion formation [133]. Trappmann *et al*. showed that, in a collagen-polymer substrate system, PDMS substrate stiffness did not affect epidermal or mesenchymal stem cell fate, while the stiffness of PAAm did. It was revealed that in PAAm changes in porosity with stiffness led to changes in collagen anchoring points, which in turn affected differentiation [134]. Also, specific integrins have been identified which interact with substrate microstructure to affect cellular behavior, including the α2β1 integrin during osteoblastic differentiation on titanium substrates [135]. More studies such as these, focusing on the mESC-fibrin system, will be needed to extract how the aforementioned microstructural features affect differentiation patterning.

3.5 CONCLUSIONS

In this study, we have developed a systems level modeling approach to investigate the contributions of various extracellular microstructural cues towards the differentiation patterning of mESC. A fibrin substrate amenable to fiber and elasticity manipulation was employed together with an integrated experimental and mathematical analysis. The combinatorial treatment and statistical analysis of the complex feature space allowed for investigation of the relative influence of individual fibrous features on ESC differentiation without the need for one-at-a time variable perturbations or large data sets. Interestingly, it was found that in this system of spontaneous mESC differentiation on fibrin gel substrates, the correlation between fibrin stiffness and gene expression was relatively weak. On the other hand, the correlation between gene expression patterning and microstructural features was strong, with fiber alignment being the most influential feature. These features preferably induced differentiation of mESC to endodermal lineage, with the majority of significant correlations involving the endoderm genes *FOXA2* and *AFP*.

4.0 STOCHASTIC POPULATION MODEL OF CELL CYCLE TRANSITION IN HESC DURING SELF-RENEWAL AND DIFFERENTIATION

4.1 INTRODUCTION

Chapter 3 focused on how different substrate cues affect end-point differentiation. Changing substrate fabrication conditions led to changing of the microstructural features, and differences in these features in turn lead to differences in gene expression. This gene expression was obtained for each substrate fabrication condition through PCR, and therefore represented population averaged information. However, even in a given substrate (or, more generally, in a given culture condition), not all cells behave the same, as is apparent with experimental data which does not rely on population averaged information, but which reports single cell behavior. This amounts to variability from cell to cell, with a resulting heterogeneous cellular population. This variability can have consequences on population behavior, and can often dictate system dynamics.

This cellular variability is especially evident in the cell cycle. The cell cycle is the process in which a cell prepares for and executes replication of itself [136]. The process begins with one daughter cell, and if conditions are such that division is supported, and if all cellular functions act properly, ends with two daughter cells with the same genomic material. ESC have a unique cell cycle structure, with short doubling times, mainly due to their abbreviated G1 phase

[38, 39]. In contrast, somatic cells have a much longer doubling time with the majority of the cell cycle spent in the G1 phase. When ESC differentiate, this G1 phase elongates, resulting in an overall longer doubling time [40, 41] and slower propagation. This G1 phase is also important in governing cell fate, as studies have shown it is involved in cellular decision making [137, 138]. An important aspect of this system is therefore the coupling of pluripotency to cell cycle. It is now known that this coupling is very important, and in addition to differentiation affecting G1 length, changes to the cell cycle, say from changes to the cyclins and CDKs, affect differentiation status [139].

Our objective in this chapter is to represent the population dynamics of cell cycle transition in human ESC during self-renewal and differentiation, and to gain mechanistic information on the process. To describe this behavior, we utilize an automaton model which tracks autonomous cells through the cell cycle. Through this we determine phase residence time distributions of hESC and a more mature pancreatic progenitor cell phenotype. In this way, we can judge how variability in the cell cycle changes upon differentiation. To describe this transition, we focus on the G1 phase, as this phase seems to be the most crucial in guiding hESC behavior. We develop a cellular ensemble model, for which we integrate both a rule-based platform and a simple system of ordinary differential equations governing the G1 phase into the cellular automata. In this way we describe G1 lengthening at the single-cell level and how this behavior is manifested at the population level. Through this approach, we quantitatively represent the cell cycle dynamics of hESC and hESC-derived pancreatic progenitors, and elucidate the mechanisms of cell cycle changes during the process of differentiation.

4.2 METHODS

4.2.1 Cell culture and differentiation

H1 hESC were cultured in feeder free conditions on Matrigel™ (hESC qualified matrix) in mTesr1® media. Cells were passaged every 6-8 days. In differentiation experiments, mTesr1® media was replaced at day 4 or 5 of culture with differentiation media. To direct towards definitive endoderm, DMEMF12, 0.2% BSA, and 1x B27® (base media) was supplemented with 100 ng/mL activin A and 25 ng/mL wnt3a and replaced daily for four days. Cells were then induced to a pancreatic progenitor phenotype with 0.2 uM cyclopamine in base media for two days, followed by another two days of cyclopamine and 2 uM retinoic acid, with media changes every day.

4.2.2 Cell cycle synchronization, flow cytometry, Fourier analysis, and CFSE

hESC were synchronized in the G2/M phase of the cell cycle with 200 ng/mL nocodazole for 16.5 hours. Cells were then rinsed two times with DMEM F12 and then mTesr1 was added to the wells. Directly after synchronization (t=0), and at three hour intervals, cells were harvested for cell cycle analysis by dissociation with Accutase®. After dissociation, cells were fixed in ice cold 70% ethanol and stored overnight at -20°C. Cells were then rinsed twice in PBS and incubated for 25 min at room temperature with 0.01 mg/mL/1 million cells propidium iodide (PI) and 0.2 mg/mL DNAse-free RNAse in 0.1% triton-X in PBS to stain for DNA. The cell cycle histogram was generated by analyzing the cellular population with an Accuri C6 flow cytometer, and the phase distribution was quantified by Modfit LT software.

To synchronize pancreatic progenitor cells, cells were first differentiated to a pancreatic progenitor phenotype as outlined above. These cells were then exposed to 200 ng/mL nocodazole in pancreatic progenitor media for 31 hours. This longer incubation time was chosen to allow for more slowly dividing cells to reach the G2 phase and arrest in the phase. Cells were then rinsed two times with DMEM F12 and pancreatic progenitor media was added to the wells. Several wells were immediately harvested for t=0 analysis, with temporal harvesting occurring every three hours. Dissociation, fixing, staining, and cell cycle analysis was performed as per the protocol outlined above for the undifferentiated cell synchronization experiment.

Discrete Fourier Transform (DFT) was performed on the oscillatory data achieved through synchronization to convert the time series data into the frequency domain. Data was first normalized around 0, and a spline fit to the experimental dynamics was performed to generate enough data points to be amenable to DFT. The Fourier transform was performed with the *fft* function in MATLAB. To attenuate high frequency noise, the data was run through a low pass, Butterworth filter with the *filtfilt* function, the design parameters of which determined by the *butter* function, both in MATLAB.

To capture the dynamics of differentiation, cells were induced towards endoderm and pancreatic progenitor as previously described. Cells were harvested at various times during this differentiation protocol, stained and analyzed for the cell cycle profile as detailed above.

Carboxyfluorescein succinimidyl ester (CFSE) dye was used to experimentally determine doubling times. H1 hESC were washed two times with PBS, then incubated in 1.5 ug/mL in PBS for 15 minutes at 37°C. Cells were then washed three times with DMEM F12 with 20% knock-out serum replacement (to help quench unbound CFSE), washed two times with DMEM F12, and then placed in mTesr1 media. Cells from several wells were then harvested (t=0), suspended

in PBS, and analyzed with flow cytometry. Proceeding this, cells were harvested and analyzed every 24 hours (those cells not harvested had daily media changes). CFSE flow cytometry data was then analyzed with Modfit LT software, and doubling times extracted.

4.2.3 Population model of the cell cycle

Experimental synchronization allows for the tracking of the cellular population through the individual phases of the cell cycle. To extract information from this data on phase residence times, distributions, and variability, and how this changes with differentiation, we develop a mathematical model of cell cycle population behavior and associated heterogeneity in the hESC system. In this mathematical approach we utilized an automaton model, in which we track individual cells and their progress through the cell cycle phases. This model was proposed by Altinok *et al.* [140], and we adapted it for the hESC system. The model starts off with a population of cells, with each cell in either the G1, S, or G2/M phase. We chose to combine the G2 and M phases because separating them will not add to the analysis, as the focus of the hESC system is the G1 phase. Each cell has two main attributes: the time it will stay in the phase and the time it transitioned from the previous phase (and therefore at any given time it can be determined how long the cell has been in the current phase). The former is determined stochastically from probability distributions. Each phase has a probability distribution describing its residence time, and as a cell enters this phase, the time it will stay in the phase is chosen from this distribution. These distributions are not known *a priori*, and have to be determined, along with their moments, with experimental data.

The number of cells in each phase at the beginning of the simulation is determined by experimental population data directly after synchronization (t=0). Because no additional

experimental information is known about how long the cell has been in the current phase at this time, this value is randomly chosen from a non-informative (uniform) distribution, from 0 (just entered the phase) to the total residence time chosen from the probability distribution (just about to transfer to the next phase). The simulation is then started, and the individual cells progress through the cell cycle. At any time, the number of cells in each phase can be determined, and therefore cell cycle population dynamics can be generated.

4.2.4 Parameter estimation

Because the probability distributions describing the phase residence times were not known, we employed a combinatorial approach. Each prospective model contained a specific combination of distributions (e.g. gamma for G1, normal for S and G2/M). These combinations are listed in Table 4.1. As others have postulated that the G1 phase might behave differently than other phases and is most involved in differentiation and fate decisions [41, 137, 138, 141-144], we put more emphasis on this phase. In the simulations in which a normal distribution was involved, a cut-off of 0 was enforced so that all residence times were strictly positive. For each combination, a genetic algorithm was used to estimate the moments associated with each distribution from agreement with experimental data (Figure 4.1). For instance, for combination 2 in Table 4.1, the genetic algorithm was employed to estimate the α and β parameters for the G1 gamma distribution, μ and σ for the S normal distribution, and μ and σ for the G2/M normal distribution, to maximize agreement between the simulation and experimental synchronization dynamics.

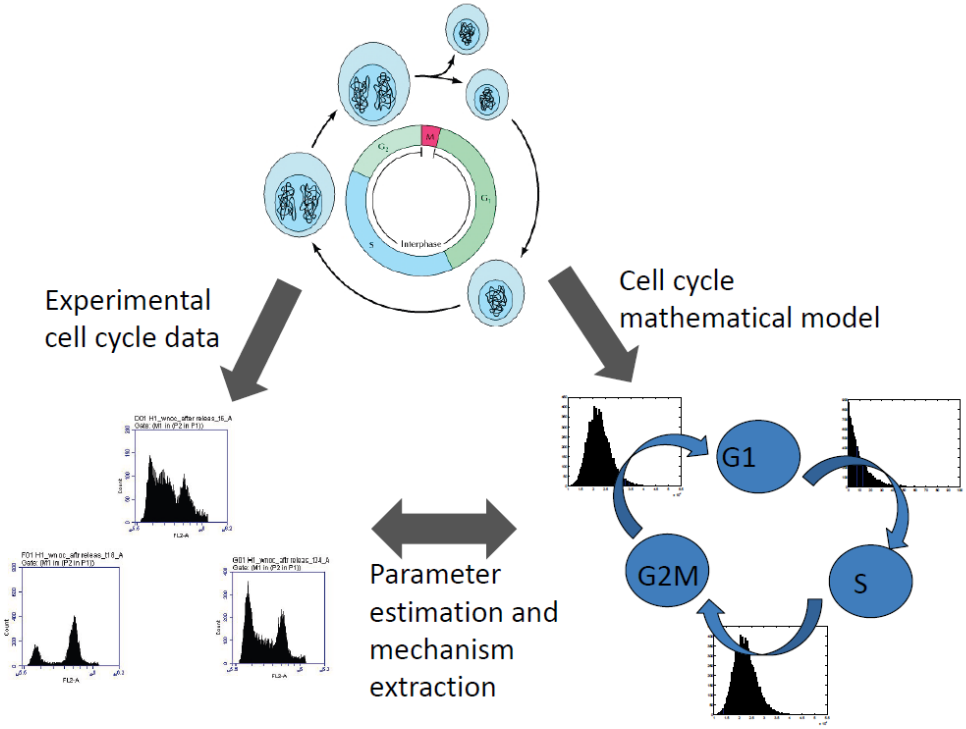

Figure 4.1. Cell cycle model development

Cell cycle structure (top) was used to develop a stochastic population model (bottom right) which utilized experimental cell cycle population data to determine cell cycle parameters and mechanisms

Table 4.1. Combinations of probability distributions describing cell cycle utilized in parameter estimation

Combination	G1 Phase	S Phase	G2/M phase
1	Normal	Normal	Normal
2	Gamma	Normal	Normal
3	Exponential	Normal	Normal

4.2.5 Cellular ensemble model

To describe G1 lengthening with differentiation, and to create a method for hypothesis testing, we developed a cellular ensemble model. This model takes the cell cycle population model and incorporates mechanisms governing G1 lengthening at the single cell level. In particular, we developed a platform in which the probability distribution governing G1 residence time is replaced by a phenomenological formulation. In the simulation, cells entering G1 phase call upon the equations, which dictate G1 residence time and how this changes with differentiation.

In parallel with this, we employed a more mechanistic approach, in which the G1 residence time and its changes are governed by a set of nonlinear ordinary differential equations (Equation 4.1-4.5 and Table A3).

$$\frac{dD}{dt} = s_D - d_D D - k_1 DI + k_2(D.I) + k_3 E(D.I) \tag{4.1}$$

$$\frac{dE}{dt} = s_E - d_E E - k_4 EI + k_5(E.I) + k_3 E(E.I) + \frac{k_6 E^H}{k_8 + E^H} + \frac{k_7(D + D.I)}{k_9 + (D + D.I)} \tag{4.2}$$

$$\frac{dI}{dt} = s_I - d_I I - k_1 DI + k_2(D.I) - k_4 EI + k_5(E.I) - k_3 EI \tag{4.3}$$

$$\frac{d(D.I)}{dt} = k_1 DI - k_2(D.I) - d_{(D.I)}(D.I) - k_3 E(D.I) \tag{4.4}$$

$$\frac{d(E.I)}{dt} = k_4 EI - k_5(E.I) - d_{(E.I)}(E.I) - k_3 E(E.I) \tag{4.5}$$

These equations describe, at the single cell level, the temporal changes of the levels of proteins which are integral to the G1 phase (cyclin D, cyclin E, inhibitors p21/p27 (noted here as 'I' for simplicity)), and their associated complexes (denoted as a '.')). Therefore, when cells enter the G1 phase, an ODE solver is called (lsode), which performs integration, and from this G1 time is determined as the time at which cyclin E switches from a low state to a high state. We adopted this ODE model from the work of Pfeuty [46] and Novak and Tyson [48], and modified it for the hESC system. As in these two studies, it is assumed that the CDKs, responsible for cell cycle transitions, are at constant high level, and therefore the transitions are governed by the temporal behavior of their catalytic partner, the cyclins. The model therefore represents this cyclin-CDK partnership in terms of just the cyclin levels. As described in these two studies, there are numerous proteins governing the G1 phase. In the current system we reduced the G1 model to a skeletal five molecule system (three proteins and two associated complexes) based on experimental evidence reported for hESC. Because of this, several simplifications needed to be made, the most crucial of these being the description of the effect of cyclin D and cyclin E on the latter's behavior. These effects come about through other "intermediate" connections and proteins, including Rb and E2F. In the absence of these proteins in the current model, we represented the interaction as Michaelis-Menten kinetics describing cyclin D's positive influence on cyclin E, and Hill type kinetics describing cyclin E's autocatalytic function. For those protein interactions not simplified, we took nominal parameters agreeing with the two aforementioned works. For those which are simplified, namely the 6^{th} and 7^{th} terms in Equation 4.2, parameters were chosen which led to behavior consistent with biological observations.

For the ensemble model, parameter estimation was accomplished using genetic algorithm with various G1 population dynamics resulting from various differentiation induction conditions (described in the section 4.3). In all models, the simulation was coded in FORTRAN 95.

4.3 RESULTS

4.3.1 Cell cycle synchrony behavior changes in hESC after differentiation to pancreatic progenitor

To assess the cell cycle dynamics of self-renewing hESC, we performed PI staining to quantify the fraction of the population in each of the three main cell cycle phases: G1, S, and G2/M. Figure 4.2(a) shows a representative histogram of the DNA stain of asynchronous H1 hESC analyzed by flow cytometry, with Figure 4.2(b) showing the phase proportions after quantification by Modfit, a program which fits raw flow cytometry DNA data to phase specific functions, with subsequent phase percentage quantification. As expected, the majority of the cells are in the S-phase, with a relatively small proportion in the G1 phase, converse to somatic cells.

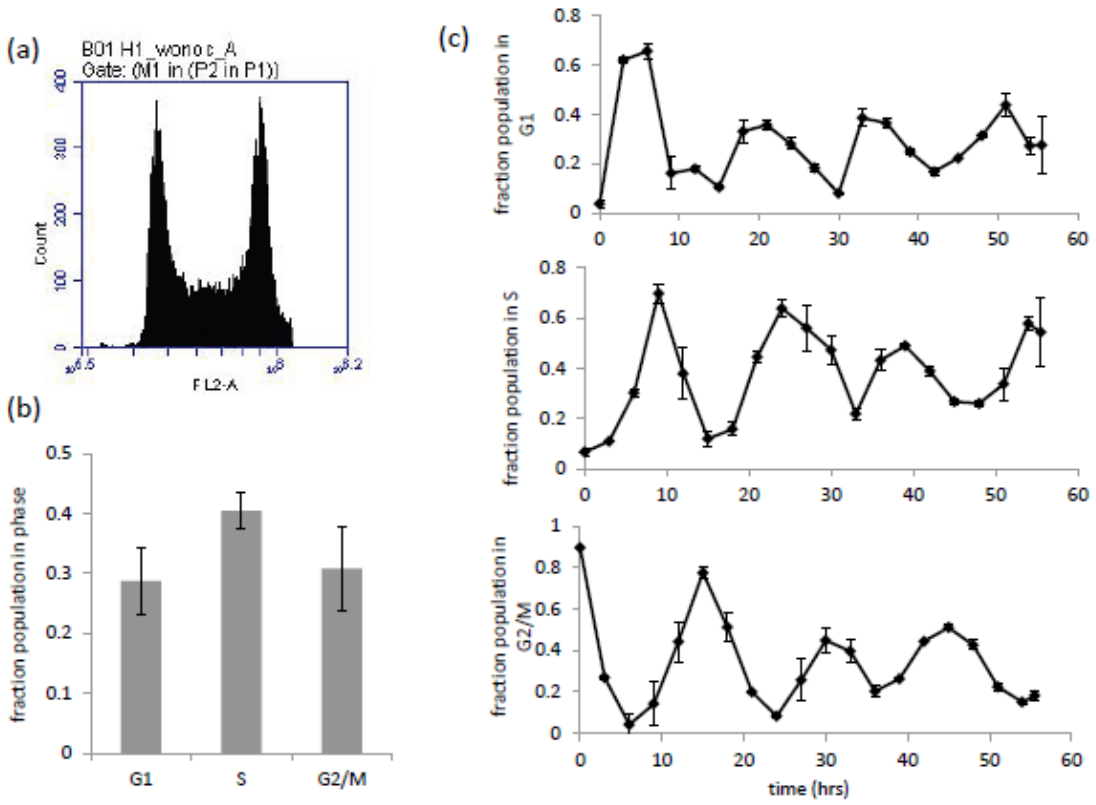

Figure 4.2. Cell cycle behavior of self-renewing hESC

(a) Representative flow cytometry histogram of DNA content for a population of H1 hESC. (b) Quantification of the cell cycle distribution, n=14. (c) Dynamics of the cell cycle distribution after release from G2/M synchronization, n=2. Error bars=std .dev.

We gained insight into the cell cycle traits of hESC, including phase transition behavior, residence time distributions and variability, by performing dynamic synchronization experiments. Using nocodazole, we were able to synchronize close to 90% of the cellular population in the G2/M phase. Upon removal of nocodazole, the cells exit the phase and re-enter the cell cycle. While initially the cell population will retain synchronization characteristics, because of variability in the system, the cells tend to lose this synchronization and revert to their asynchronous state. The variability, or lack thereof, in the system can be determined by analyzing how fast the synchronized oscillations dampen after nocodazole removal. Figure

4.2(c) shows the dynamics of the cell population upon release from synchronization, displaying how the fractions of each of the phases change after G2/M synchronization. A striking feature of these dynamics is the retention of synchronization characteristics over the time course (~3.5 cell cycles) of the experiment when compared to other cells which desynchronize more quickly. While there seems to be some initial loss of synchrony of the ESC within the first cell cycle, afterwards the oscillations do not completely dampen within 3.5 cell cycles, which is in contrast with synchronization studies of other cellular systems, which display desynchronization within three cell cycles [60, 145, 146]. This might be due to possible differences in the cell cycle and phase residence time variability between ESC and somatic cells. The cell cycle of somatic cells has been shown to be highly variable [141], which explains dramatic desynchronization. Low ESC cell cycle variability would explain the observed retention of synchronization characteristics.

To analyze if this behavior is preserved with maturation, we differentiated hESC, first to definitive endoderm, then to pancreatic progenitor cells, and assessed the cell cycle. Flow cytometry at each stage of differentiation for indicative phenotype proteins (sox17 for endoderm, pdx1 for pancreatic progenitor) confirms differentiation (Figure 4.3(a)). Figure 4.3(b) shows the cell cycle phase distribution. As expected, with differentiation to a more mature phenotype, the G1 has lengthened, as shown by a larger proportion of the population in the phase. We performed synchronization on these cells and analyzed the cell cycle dynamics of this phenotype upon release from synchronization. Exposing the pancreatic progenitor cells to 200 ng/mL nocodazole (in pancreatic progenitor base media with cyclopamine and retinoic acid) for 31 hrs lead to 51% percent synchronization in the G2/M phase, at which point the nocodazole was removed and replaced with pancreatic progenitor media. The resulting dynamics are shown in

Figure 4.3(c). In stark contrast to the dynamics of the pluripotent cells, the synchronized oscillations in the pancreatic progenitor cells dampen out quite quickly after release from synchronizaion. By 24 hrs the cell cycle distribution is close to the asynchronous distribution. This might be an indicator of increased cell cycle variability with differentiation, which we further analyzed using our developed model.

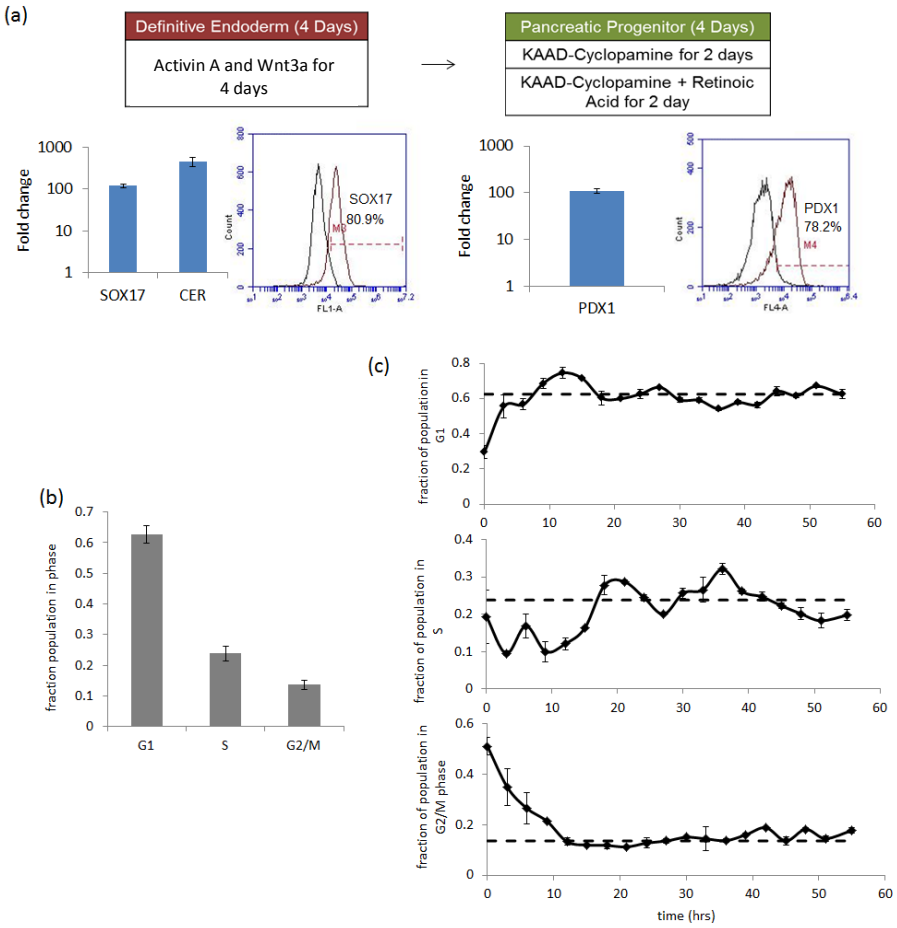

Figure 4.3. Cell cycle behavior of hESC-derived pancreatic progenitor cells

(a) Schematic of differentiation protocol, showing presence of stage specific markers at end of differentiation stage. Gene fold change with respect to undifferentiated cells, quantified by qPCR. Flow cytometry showing population positive for respective protein. Black histogram is 2ndary antibody only control. (b) Quantification of the cell cycle distribution of hESC-derived pancreatic progenitors, n=3. (c) Dynamics of the pancreatic progenitor cell cycle distribution after release from G2/M synchronization, n=2. Dashed line represents asynchronous distribution. Error bars=std. dev.

Another interesting feature of the oscillations is the irregular waves throughout the time series. If one considers the S-phase dynamics, superficial analysis of the waves seems to suggest that there are multiple wavelengths present. For instance, the distance between the two highest peaks is around 15 hrs, while the distance between the valleys at the middle and end of the time course is around 24 hrs. This is in addition to the multiple valleys at the beginning of the experiment. These features are reminiscent of multiple waves superimposed onto one another. This could happen if there were two subpopulations of cells with different cell cycle characteristics. To evaluate this, we performed a Discrete Fourier Transform (DFT) analysis on the data. This analysis transforms the data from the time domain into the frequency domain, and can therefore allow for the determination of prominent frequencies present in the data. The results of this analysis for the individual phases are shown in Figure 4.4 (a-c), and suggest that oscillations with several frequencies occur in the data set. When analyzing the frequencies in Figure 4.4(a), there exists a broad higher frequency peak at ~0.125 hr^{-1}, which corresponds to a wavelength, and in the context of the current data set cellular doubling time, of 8 hrs. This is much too short for hESC, and therefore not a realistic oscillation, and can be considered high frequency noise, which has been observed elsewhere [146, 147]. To clean this data and remove this high frequency noise, we applied a low-pass Butterworth filter to the data (Figure 4.4(d)), which results in two prominent frequency peaks of 0.0178 hr^{-1} and 0.0533 hr^{-1}, or doubling times of 56.2 hrs and 18.8 hrs, respectively (Figure 4.4(e-g)). This would indicate that proliferation of two subpopulations arising during differentiation were both retarded, but by different amounts. Together, these results suggest that while propagating hESC retain cell cycle synchrony with time, the behavior is lost with maturation towards pancreatic lineage, with this transition also bringing about an emergence of multiple populations displaying distinct phase characteristics.

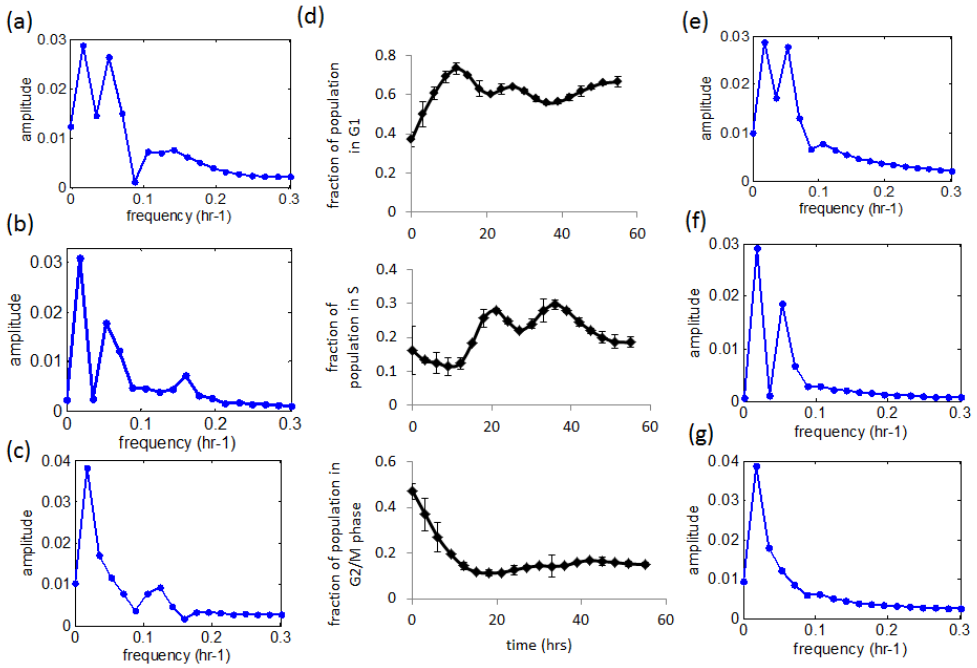

Figure 4.4. Frequencies in synchronized pancreatic progenitor cells

(a-c) Discrete Fourier transform (DFT) applied to the experimental pancreatic progenitor synchronization data of the G1 (a), S (b), and G2/M (c) phases. (d) Experimental dynamics after noise removal with a Butterworth filter. (e-g) DFT analysis applied to G1 (e), S (f), and G2/M (g) phase data after noise removal

4.3.2 Stochastic population model extracts single-cell information from population dynamics of differentiating hESC

We first developed a stochastic population model to track the dynamics of the cell cycle after relaxation from synchronization. An advantage of this type of model is that it allows for tracking the emergent cell population dynamics arising from single cell variability and features. Each cell progresses through each phase of the cell cycle following probabilistic rules. The basis of the cell cycle model utilized herein is that the cells are in a state (cell cycle phase), and the time in that state is stochastically chosen from a probability distribution. In particular, we adopted the approach by Altinok *et et al*. [140], wherein discrete, autonomous cells are tracked

with time, thereby simulating the behavior of a cellular population. The variables in the model are the probability distributions describing the residence times of the phases (normal, exponential, etc.), and the moments associated with each residence time (μ and σ for normal, λ for exponential, etc.).

We applied the developed model to the measured dynamics of self-renewing hESCs. Since the type of distribution describing each phase was not known *a priori*, we employed a combinatorial approach to determine these distributions. For each combination of distributions, listed in Table 4.1, genetic algorithm was used to estimate the moments associated with the phase residence time distributions from experimental synchronization data. Akaike information criterion (AIC) was used to determine which combination of phase distributions were optimal. AIC, based on information theory, attempts to determine the relative quality of a model, and can therefore be utilized to select an optimal model/approach from a set of alternatives by minimizing information loss. AIC weighs the model fit against model complexity, namely the number of free parameters [148]. There are several ways of reporting AIC for a set of alternatives. To compare cell cycle model alternatives, we utilize AIC weights, which can be thought of as the probability that a given model is the best model of a set [149] (Equations 4.6 and 4.7)

$$AIC_i = n\ln\left(\frac{SSE}{n}\right) + 2k \tag{4.6}$$

$$weight_i = \frac{e^{-0.5(AIC_i - AIC_{\min})}}{\sum_i e^{-0.5(AIC_i - AIC_{\min})}} \tag{4.7}$$

where n is the number of data points used to train the model, SSE is the sum of the squares of the residuals, k is the number of model parameters, and AIC_{min} is the minimum AIC out of the set. In practice, if one model alternative's weight is not clearly higher than others', a single model cannot be conclusively selected. In the current approach, we have determined the AIC weight for each of the three distribution combinations in Table 4.1, and are looking for a combination which has a weight that is much better than the rest, which would suggest an optimal combination. While the simulation is stochastic, we took one instance of the simulation (one Monte Carlo run) for each combination to determine the AIC weight. This analysis determined the Akaike weights for distributions 1, 2, 3 (Table 4.1) as 0.380, 0.591, and 0.029, respectively. This metric shows that while combination 3 (exponential describing G1 phase) performed poorly, both combinations 1 and 2 (normal and gamma distributions, respectively, describing G1 phase) yield relatively large probabilities of being the best candidate model. Therefore, there is an amount of redundancy in the model. However, the phase means and standard deviations predicted from both of these combinations are almost identical, and the gamma distribution resulting from the predicted α and β parameters is roughly Gaussian in nature. We can therefore say that the G1 phase is normally distributed. The model dynamics and agreement to the experimental dynamics, as well as the predictions of the normal distribution parameters for each phase, are shown in Figure 4.5(a) and Table 4.2(a), respectively. As shown, the model does an excellent job at capturing the oscillations in cell cycle distribution associated with synchronization and release. The predicted G1 phase residence time is short, 3.2 hrs, as expected for hESC. The variability of these phases is quite low as well, something which may not be expected in somatic cells, especially for the G1 phase.

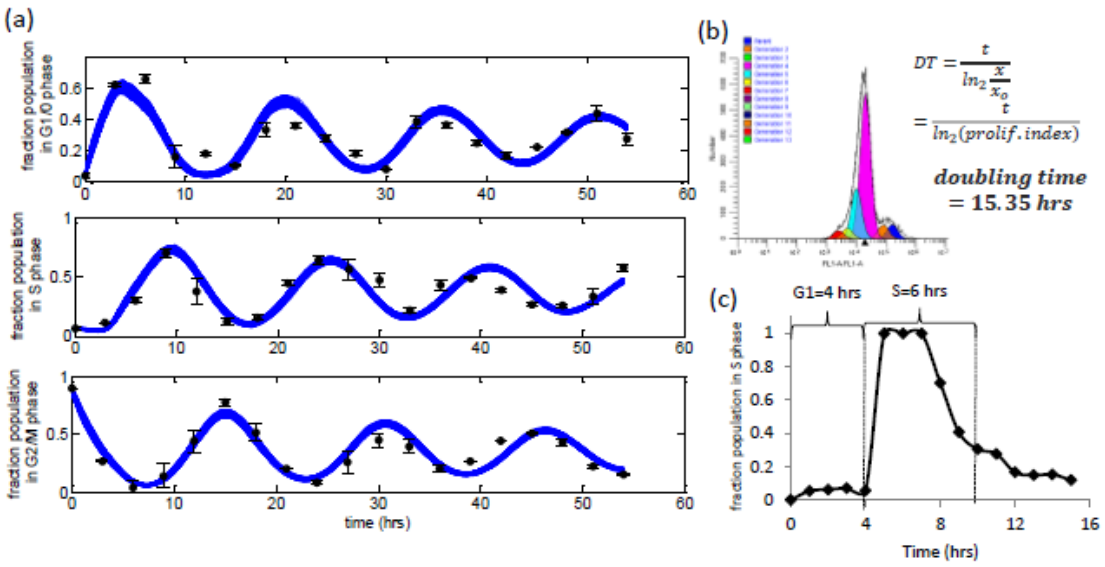

Figure 4.5. Cell cycle model predictions of the synchrony behavior of undifferentiated hESC

(a) Simulated dynamics (blue band, 1000 stochastic runs), fitted to experimental data (black points, n=2, error bar=std. dev.) for the three phases. (b) Validation of total cell cycle time with CFSE staining for doubling time quantification. (c) Validation of G1 and S residence times with small time step DNA staining quantifying S-phase fraction after G2/M synchronization.

Table 4.2. Predicted cell cycle parameters from synchronization experiments

Model predictions of the normal distribution parameters for each phase of (a) Self-nenewing hESC and (b) hESC-derived Pancreatic progenitor

(a)	Mean (hrs)	Std. Dev. (hrs)
G1	3.24	0.58
S	5.61	1.76
G2/M	6.67	0.74

(b)	Subpopulation 1=61.2%		Subpopulation 2=38.8%	
	Mean (hrs)	Std. Dev. (hrs)	Mean (hrs)	Std. Dev. (hrs)
G1	18.6	8.6	15.6	2.77
S	15.6	5.3	4.7	0.57
G2/M	22	10	1.2	0.24

To validate our model and the predicted results, we experimentally measured the doubling time and G1 and S phase residence times using flow cytometry on a completely separate cell culture (Figure 4.5 (b,c)). Using CFSE analysis to quantify doubling time yielded an average doubling time of 15.4 hrs, in excellent agreement with model predictions, achieved by summing the predicted individual mean residence times. To measure phase residence times, we again performed synchronization on the hESC cell culture, but analyzed the phase distributions at small time increments (1 hr) for better resolution. Shown in Figure 4.5(c) is the S phase population fraction. Because at t=0 the population has been synchronized in the G2/M phase, there are few cells in the S phase. The time at which this proportion jumps (4 hrs) corresponds to the G1-phase length. Additionally, the S-phase time corresponds to when the S-phase proportion substantially decreases (10 hrs), minus the G1 time (S-phase = 6 hrs). Again, this is in excellent agreement with our model predictions, validating the accuracy of the model and estimated parameters.

We then applied our population model to the analysis of pancreatic progenitor cells. Because the DFT analysis suggested that two subpopulations, with two distinct frequencies, might contribute to the overall oscillations, in the population model we simulated two populations of cells, the phases of which being described by normal distributions (as predicted by the self-renewing system), each with their own set of cell cycle phase characteristics. These predicted dynamics as compared to experimental data, as well as a visualization of the individual subpopulation dynamics which contribute to this behavior, are shown in Figure 4.6. Table 4.2(b), shows the predicted phase parameters. The model predicted that the split between the two populations is ~61/39%, with the larger subpopulation displaying almost proliferation arrested conditions, as the total cell cycle time is approximately the total experimental time. The

variability in this subpopulation is also very high, which would lead to quick reversion to an asynchronous state, as demonstrated by the green curves in Figure 4.6(b). The other subpopulation shows behavior which suggests it is at a more immature state, as the cells are still doubling in a relatively short amount of time (~18 hrs); however, G1 has lengthened substantially from undifferentiated cells, and therefore the population has matured somewhat. Also, the variability associated with this subpopulation is lower than the first, suggesting that variability increases with differentiation. Together, these results demonstrate the utility of the population model in capturing single-cell level dynamics from cell population information. Encouragingly, the model exhibited excellent performance even with a heterogeneous differentiated population and could extract sub-population dynamics.

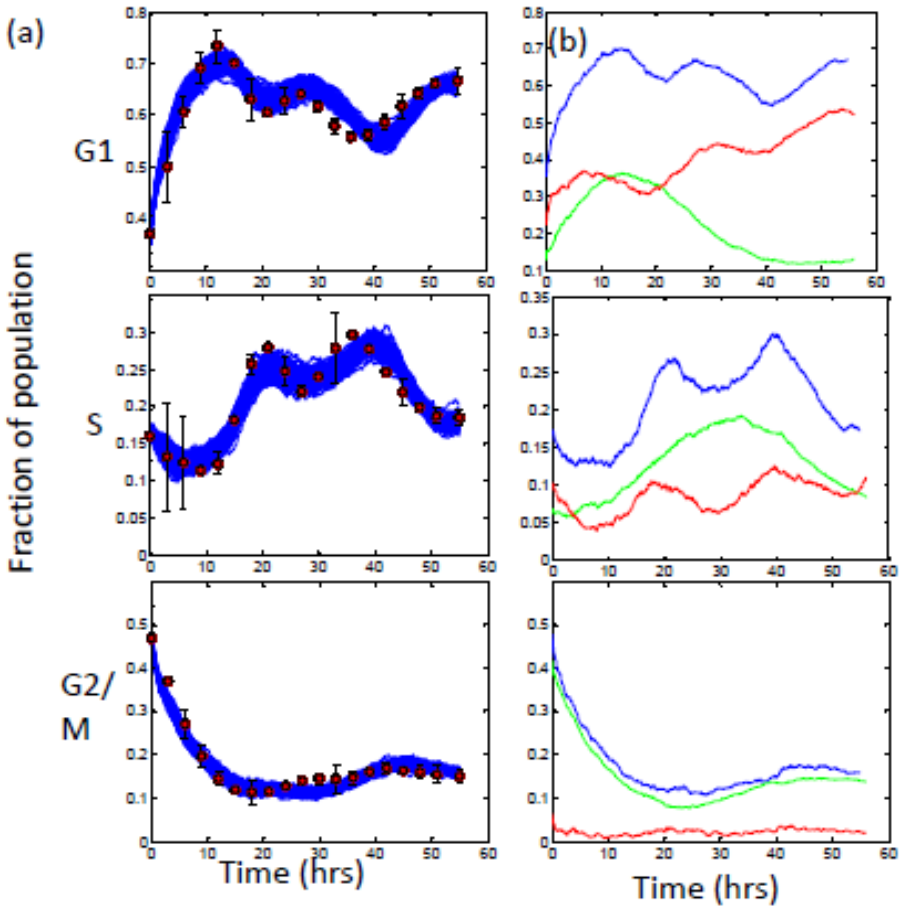

Figure 4.6. Cell cycle model predictions of the synchrony behavior of hESC-derived pancreatic progenitor cells

(a) Combined simulation dynamics of both subpopulations and their contributions towards cell cycle phase oscillations (blue band, 1000 stochastic runs) as compared to experimental data (black points, n=2, error bar=std. dev.). (b) Simulated dynamics of individual populations. Blue: total, Green: subpopulation 1; Red: subpopulation 2.

4.3.3 Mechanisms of G1 lengthening during differentiation revealed by cellular ensemble model

The population model when applied to the self-renewal and differentiated hESC could clearly capture (i) the short G1 residence time in self-renewing hESC, (ii) the lengthening of G1 residence time with differentiation, and (iii) enhanced variability and population heterogeneity with differentiation. Next, we wanted to capture the transition from self-renewal to differentiation, i.e. the cell cycle dynamics during the process of differentiation.

To capture and explain these trends, we needed to account for how G1 is lengthening in the interim during differentiation. The proposed model achieves this by capturing single cell behavior, and how this behavior manifests itself at the population level. Because the developed model tracks cellular automata, the model is directly amendable to this, and information on how a single cell's G1 phase lengthens with time can be incorporated into each individual cell. The G1 length is governed by a complex network of proteins controlling cell cycle progression. As a first step we wanted to determine a phenomenological model capturing the quantitative dynamics of G1 lengthening with time. Several alternate dependencies between G1 lengthening and time, along with their parameters, were considered. Furthermore, various alternatives on how and when a cell is primed for differentiation, thereby enabling G1 lengthening, were also considered.

Elucidating the actual behavior of single cell lengthening, and deciding on which alternate mechanism in our model is prominent, requires model comparison to experimental data. We tested our model against a system which allows for real time quantification of G1 cells, the FUCCI reporter system. In this system, an oscillating protein, specific for the G1 phase, was tagged with a red fluorescent probe. In this way, cells in the G1 phase, and their time in the phase, can be determined very precisely in real time with viable cells, without the need to fix and

do subsequent DNA staining. Recently Calder *et al*. [41] has reported the performance of FUCCI reporter system in hESC. Using this reporter they have tracked the G1 characteristics of cells during pluripotency and initial differentiation with time lapse microscopy and image analysis. With exposure to differentiation agents like DMSO, a gradual transition was observed in the cell cycle dynamics, with fraction of G1 positive cells increasing with DMSO concentration (Figure 4.7(a-c); these experimental G1 population dynamics, in addition to those in Figure 4.7(d-i), Figure 4.10 and 4.14(a-c), were generated by Calder *et al*. [41], published by Mary Ann Liebert, Inc., New Rochelle, NY). Further, the dynamics of the transition displays interesting non-intuitive behavior, which can be explained by our ensemble model as we will show herein. Several features of the experimental dynamics deserve attention. First, in the majority of the samples, there is a decrease in the population in G1 towards the end of the experiment (green circles, Figure 4.7(b,c)). Another striking feature is the oscillations and "step" behavior in the dynamics. The G1 fraction seems to increase as jumps rather than a linear increase with time. This behavior is most evident in the initial (up to 1.75 days) and final (after 2.75 days) dynamics of the 0.75% DMSO sample (purple circles, Figure 4.7(b)). At around 1 day in the 1% DMSO sample, a step-like dynamics is prominent (black circle, Figure 4.7(c)), wherein the G1 fraction increases, then remains constant for approximately 12 hours, then proceeds to increase again. Another characteristic of the DMSO samples is the delay in the effect on G1 (red circles, Figure 4.7(a,b)). The G1 percentage does not increase immediately but remains at undifferentiated levels for a time, and then proceeds to increase. This delay is not consistent throughout all samples, but seems to be dependent on DMSO concentration.

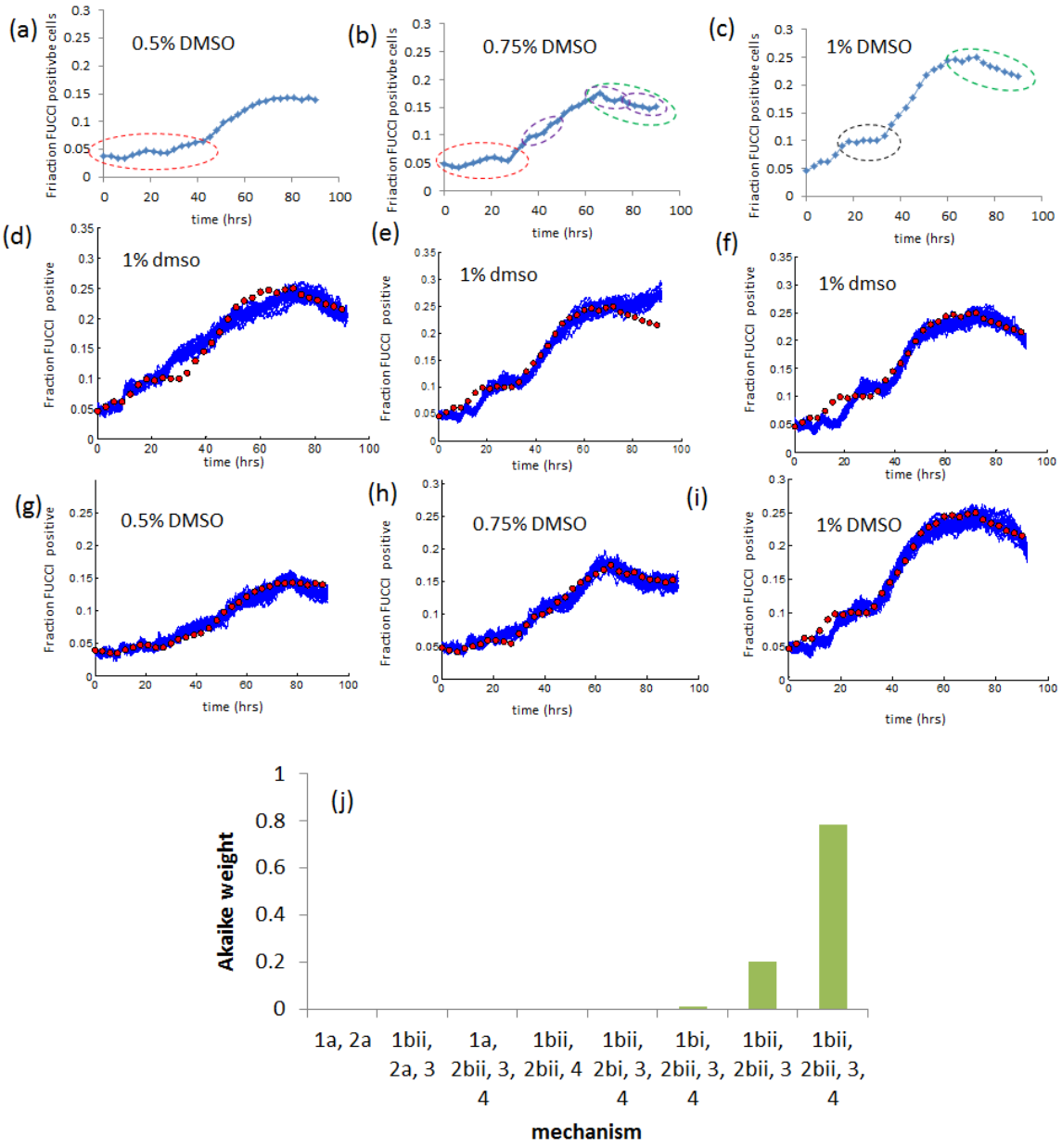

Figure 4.7. Application of ensemble model to induced differentiation with DMSO

(a-c) Experimental population dynamics of FUCCI positive cells upon exposure of hESC to 0.5% (a), 0.75% (b), and 1% (c) DMSO. Reproduced from Calder *et al.*[41]. (d-f) Simulated G1 dynamics of 1% DMSO using mechanisms 1.a, 2.b.ii, 3. (d), 1.b.ii, 2.b.ii (e), and 1.b.i, 2.b.ii, 3. (f). (g-i) Dynamics of optimal model alternative (mechanism 1bii, 2bii, 3, 4) for DMSO concentrations of 0.5%, 0.75%, and 1%, respectively (blue band, 20 stochastic simulations). (j) Akaike weights for each of the mechanisms, averaged over all DMSO concentrations. See main text for description of mechanisms.

These experimental dynamics were utilized for model hypothesis testing. It is worth mentioning that the values reported in Figure 4.7 (a-c) are fraction of population FUCCI positive. This fraction is lower than the actual G1 fraction because of the time it takes for the FUCCI reporter to mature [41], which is evident when comparing undifferentiated values in Figure 4.7(a-c) (t=0) and Figure 4.2(b). This FUCCI maturation time was taken into account in the model, and henceforth, subsequent model dynamics when compared to the experimental dynamics of Calder *et al*. are reporting the fraction of cells in G1 which have surpassed this maturation time, with figures labeled accordingly.

To explain cell cycle behavior, eight alternate G1 lengthening models were considered, which include different mechanisms which we wanted to test. In the proceeding model alternatives, we refer to differentiation, in the current context, as the state at which the G1 phase lengthens. The following are specific behaviors in the stem cell system which were modeled, with plausible mechanistic alternatives to govern each behavior:

1. Proportion of cellular population committed to differentiation
    a. Cells committed to G1 lengthening are determined from the onset of differentiation, with the number of differentiated cells only changing through mitosis
    b. Differentiation commitment is a dynamic process, with a cell probabilistically choosing whether to differentiate each time it passes through the G1 phase
        i. If a cell is committed to differentiate, G1 lengthening starts immediately upon entering the G1 phase
        ii. A cell, upon deciding to differentiate in the G1 phase, is first "primed", with lengthening only occurring upon the next cell cycle
2. Nature of G1 lengthening

    a. Once a cell is committed to differentiate, the G1 phase lengthens by a specific amount, the mean and variance of which is dictated by the differentiation agent
    b. Once a cell is committed to differentiate, the G1 phase gradually lengthens with time
        i. Lengthening is a linear function of time
        ii. Lengthening is an exponential function of time
3. Reduction of differentiation agent efficacy: Does differentiation and G1 lengthening stop at a specific time, depending on the differentiation agent?
4. Maximum time on G1. Is there an existence of a maximum threshold of G1 residence time beyond which the phase will not increase?

These alternatives were hypothesized to be possible mechanisms to explain cell cycle population behavior. Importantly, on the basis of experimentation alone it is not intuitive which of these mechanisms are governing cell cycle behavior, and it was the goal to elucidate this information with the ensemble model.

Each of these mechanisms was incorporated into the individual cellular automata of the population model, replacing the G1 probability distributions. These various models were simulated and compared to the experimental dynamics. The simulated dynamics for these model alternatives are shown in Figure 4.7(d-i) and Figure A1 and A2. As shown, the G1 dynamics are sensitive to the model structure, with different alternate mechanisms yielding drastically different population behavior. The majority of the mechanisms do not capture the observed experimental dynamics. Dynamics of alternatives without a dynamic change in differentiated cells (mechanism 1a) do a particularly poor job of describing the population behavior, as do mechanisms which have G1 lengthening by a specific amount. This suggests that cells commit to differentiation probabilistically with time, and once differentiated, have G1 phases which lengthen gradually with time (mechanisms 1b and 2b, respectively). Further insight can be obtained when analyzing the 1% DMSO dynamics (Figure 4.7 (d-f, i). A stop in differentiation and G1 lengthening

(mechanism 3) is necessary to explain the reduction in G1 population fraction. Furthermore, this data suggest that “G1 priming”, the mechanism by which cell choose to differentiate in the G1 phase, but only start G1 lengthening in the subsequent cell cycle (mechanism 1.b.ii), is necessary to explain the step behavior at ~20 hrs.

Through this comparison and agreement with experimental data, the best alternate mechanism was found to be when differentiation and G1 lengthening was not instantaneous; instead the cells were primed for differentiation with subsequent G1 lengthening after a lag. This was incorporated in the model by probabilistically deciding which cell will differentiate. The G1 length of the chosen cell, however, will not be affected immediately. Instead the cell is considered primed for differentiation, and the G1 will begin to lengthen starting the next cell cycle. Interestingly, prior literature also reports a delay between cell cycle changes and pluripotency perturbations [150]. In addition, an exponential increase in G1 with time was found to best explain the experimental dynamics. Such lengthening, however, is not indefinite and it ceases beyond a certain G1 length. Biologically, this may occur if, after a certain time, the efficacy of the differentiation factors diminishes. This model selection can be quantified by Akaike information criterion. The Akaike weights of each of the mechanism combinations is shown in Figure 4.7(j), and can represent the probability that the model alternative is the best model out of all of the model candidates. The optimal model outlined above displays the largest Akaike weight, further suggesting that the mechanisms outlined therein are governing the cell cycle behavior. The simulated dynamics of this optimal model alternative and the agreement to experimental data are show in Figure 4.7(g-i). The formulism is outlined in Equation 4.8 and pseudo-code in Figure 4.8(a) with descriptions of variables in Table A4.

$$\tau_{G1} = \min\{\alpha, \beta\}$$

where

α=maximum G1 time (parameter)

$$\beta = \begin{cases} \tau_{G1,0} & \text{if cell is undifferentiated} \\ \tau_{G1,0} e^{\gamma(T-T^*)} & \text{if cell is differentiated} \end{cases} \quad (4.8)$$

$\tau_{G1,0} \sim N(\mu_{G1,undiff}, \sigma_{G1,undiff})$ From synchronization analysis

$\gamma \sim N(\mu_\gamma, \sigma_\gamma)$ (parameter)

T=min{t,tstop} (tstop: parameter)

$T^*=t_{prime}$

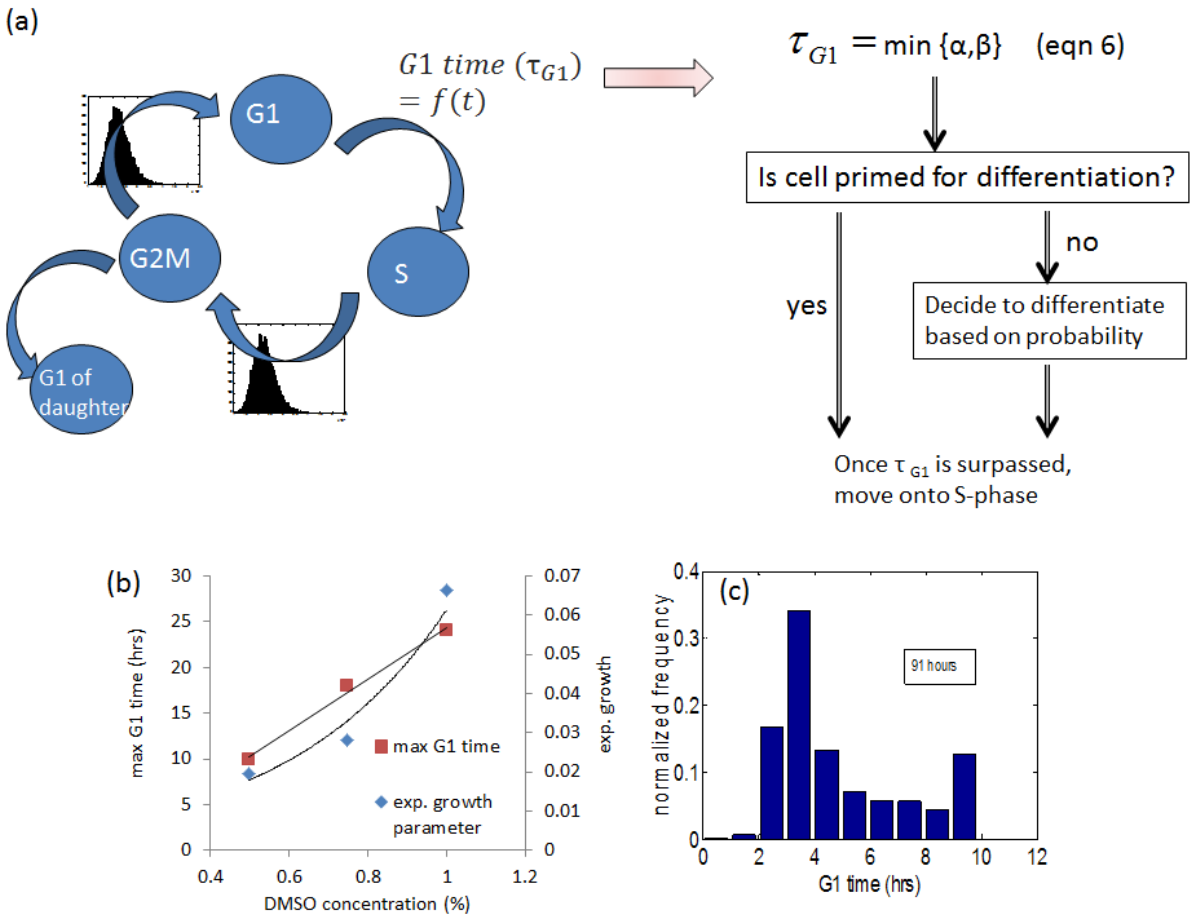

Figure 4.8. Optimal cellular ensemble model

(a) Schematic and psuedo-code of optimal model alternative for the cellular ensemble. (b) Dependency of G1 lengthening parameters (Equation 4.8) on DMSO concentration. (c) Snapshot of predicted G1 residence time at 91 hours after exposure to 0.5% DMSO.

As shown in Figure 4.7, the proposed mechanism captures the experimental dynamics to a high degree. Comparison of these simulated dynamics to other model alternatives can give an idea of the causes of the interesting features outlined above. For instance, the decrease in G1 percentage seems to arise because of a dilution effect: undifferentiated cells continue to proliferate at a faster rate than the differentiated cells, thereby increasing in number and bringing the average length of G1 (and therefore the G1 percentage) down. However this behavior is observed only when we limit the effect of differentiating agent by 'tstop'; as exclusion of this feature (not including 'tstop' in the formulation), does not yield this behavior. Another conclusion is that the oscillations seem to arise from differentiation priming only in the G1 phase. The prominent step increase in G1 percentage (1% DMSO) was found to be resulting from our hypothesis of 'priming' the cells prior to increasing the length of G1 in the subsequent cell cycle. Furthermore, the probabilistic nature of differentiation priming and the form of G1 lengthening (exponential) could adequately capture the initial delay observed in the increase in G1, and did not require an explicit delay parameter in the model. Hence even this simplistic phenomenological model could adequately explain the observed non-intuitive cell cycle dynamics. The sensitivity of the simulated dynamics to the model rules strengthens the usefulness of this platform for hypothesis testing.

In addition to elucidating the mechanisms of G1 lengthening by analyzing model structure, information can be obtained by analyzing the optimal parameter set. These values, for the DMSO experiment, and how they change with DMSO concentration, are shown in Figure 4.8(b). A clear relationship is apparent between DMSO concentration and both the G1 exponential length parameter (γ in Equation 4.8) and the maximum G1 time (α in Equation 4.8). It is interesting to note that the probability of differentiation does not change with DMSO

concentration (18.6±0.1%). Also, the time at which differentiation stops is consistent between all concentrations (63.0±5.1 hrs), and might suggest the time at which DMSO loses its efficacy.

Next we evaluated how the population distribution of G1 evolves over time, as G1 is gradually lengthening. Figure 4.9(a-c) shows the G1 residence time distribution and how it changes with differentiation for the DMSO system. In this system, the G1 phase lengthening is gradual, which causes the right tail of the G1 distribution to extend, leading to a more skewed distribution, resembling a log-normal or gamma distribution (Figure 4.8(c)). This is similar to how other studies have described cell cycle distributions [53, 141, 151] in more mature cellular systems, suggesting that as hESC differentiate, their distributions starts to resemble that of somatic cells. Together, these results show that the developed ensemble model can elucidate the mechanisms of G1 lengthening and predict pertinent cell cycle dynamic behavior.

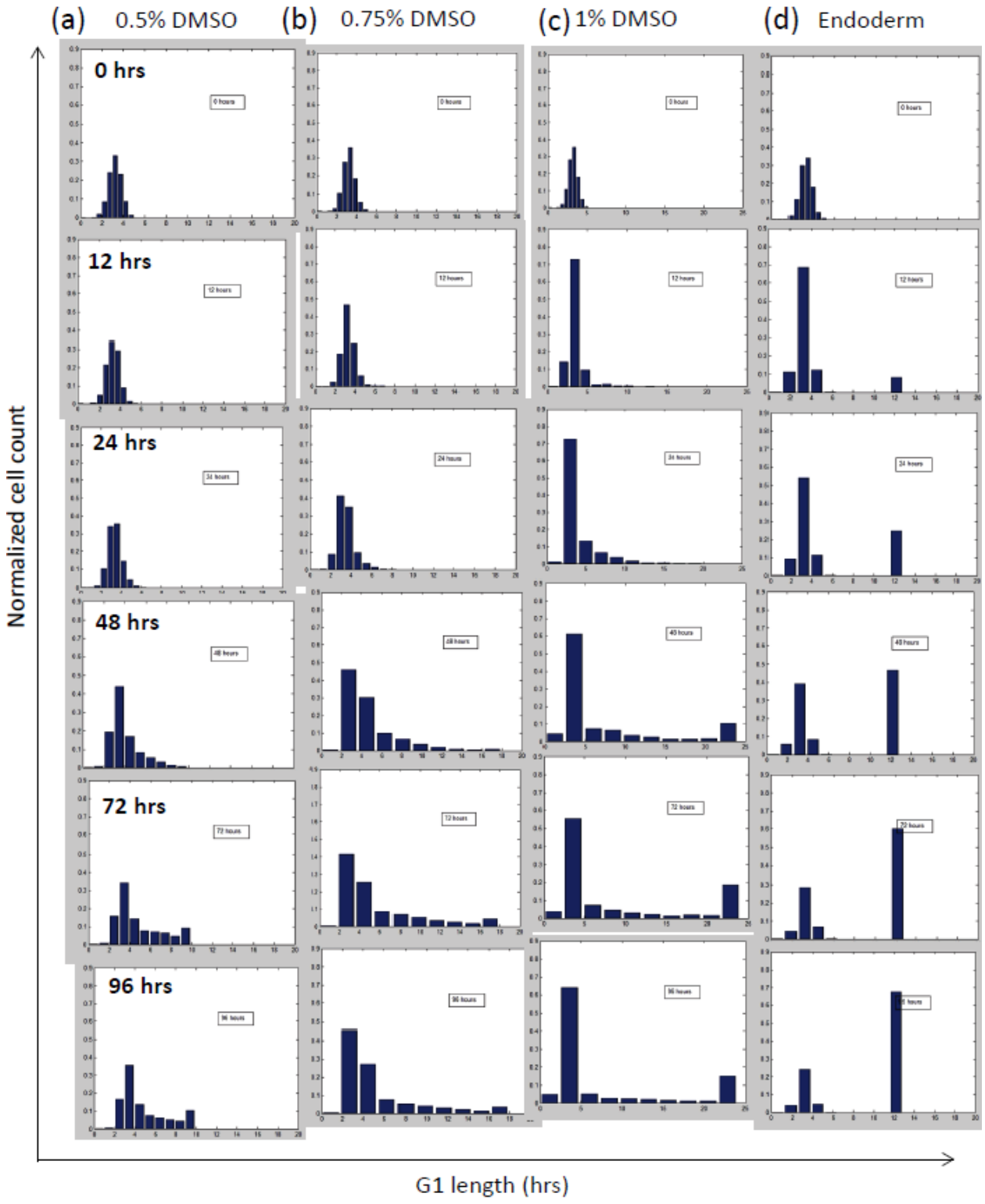

Figure 4.9. Dynamic G1 residence time

As a form of validation, this model was next applied to capture the G1 dynamics of differentiating hESC, when induced with other small molecules [41]. The model rules were kept invariant while the model parameters were re-estimated. As shown in Figure 4.10, the model does an excellent job in capturing the unique and nonlinear dynamics of G1 changes during exposure to various differentiation agents and concentrations, further validating the model rules in accurately representing the dynamics of cell cycle transition during hESC differentiation.

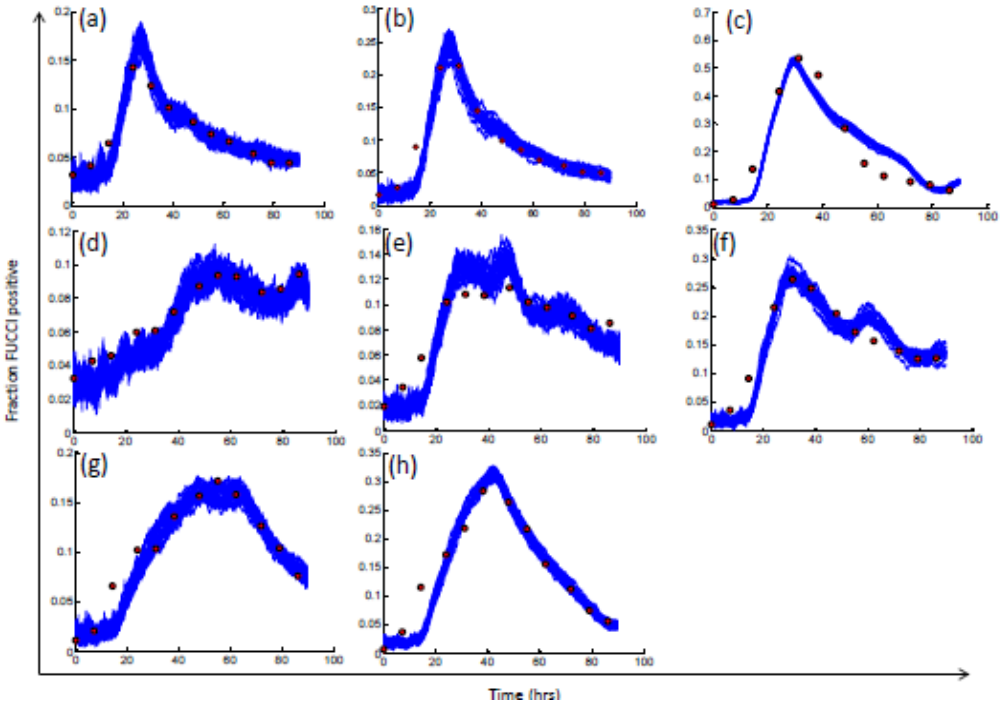

Figure 4.10. Application of ensemble model to induced differentiation with various small molecules

Model predicted dynamics (blue band, 20 stochastic simulations) compared to experimental data (Calder *et al* [41]). (a-c) 10, 20 , and 40 uM RRD, respectively; (d-f) 10, 20, and 40 uM LY, respectively; (g,h) 1.5, 3 mM HMBA, respectively

4.3.4 Oscillatory dynamics and emergence of two separate populations during endoderm differentiation explained by ensemble model

We next applied our ensemble model to capture the G1 transition during directed differentiation towards definitive endoderm lineage. The experimental dynamics of the cell cycle transition during directed differentiation towards the endoderm and pancreatic progenitor lineages are displayed in Figure 4.11(a-c). In this system, hESC were guided towards endoderm fate by exposure to proteins Wnt3a and Activin A, with subsequent exposure to cyclopamine and retinoic acid to promote pancreatic progenitor commitment via sonic hedgehog inhibition [152]. With directed differentiation, the cells mature, with a gradual lengthening of G1. This is in contrast to DMSO, which directly promotes G1 lengthening by modulating Rb protein activity [153]. When analyzing the cell cycle dynamics during endoderm differentiation, we did not utilize the FUCCI reporter, but rather performed DNA staining with flow cytometry to quantify the cell cycle distribution. This difference in G1 quantification techniques is the reason for the large discrepancy in G1 fraction for undifferentiated cells (t=0) between Figure 4.11(a) and Figure 4.7(a-c): the FUCCI reporter takes time to mature, which is manifested as an underestimate of the G1 length [41] which in turn results in a lower G1 fraction. This is not encountered with DNA stains, and therefore the G1 fraction is considerably higher. The definitive endoderm/pancreatic progenitor system displays an increase in the G1 phase proportion, with most of the change happening within the first four days (definitive endoderm induction). After this point, the G1 phase does increase, but at a slower, more gradual rate. An interesting feature of these dynamics is the oscillatory behavior of this increase, especially during the onset of differentiation induction.

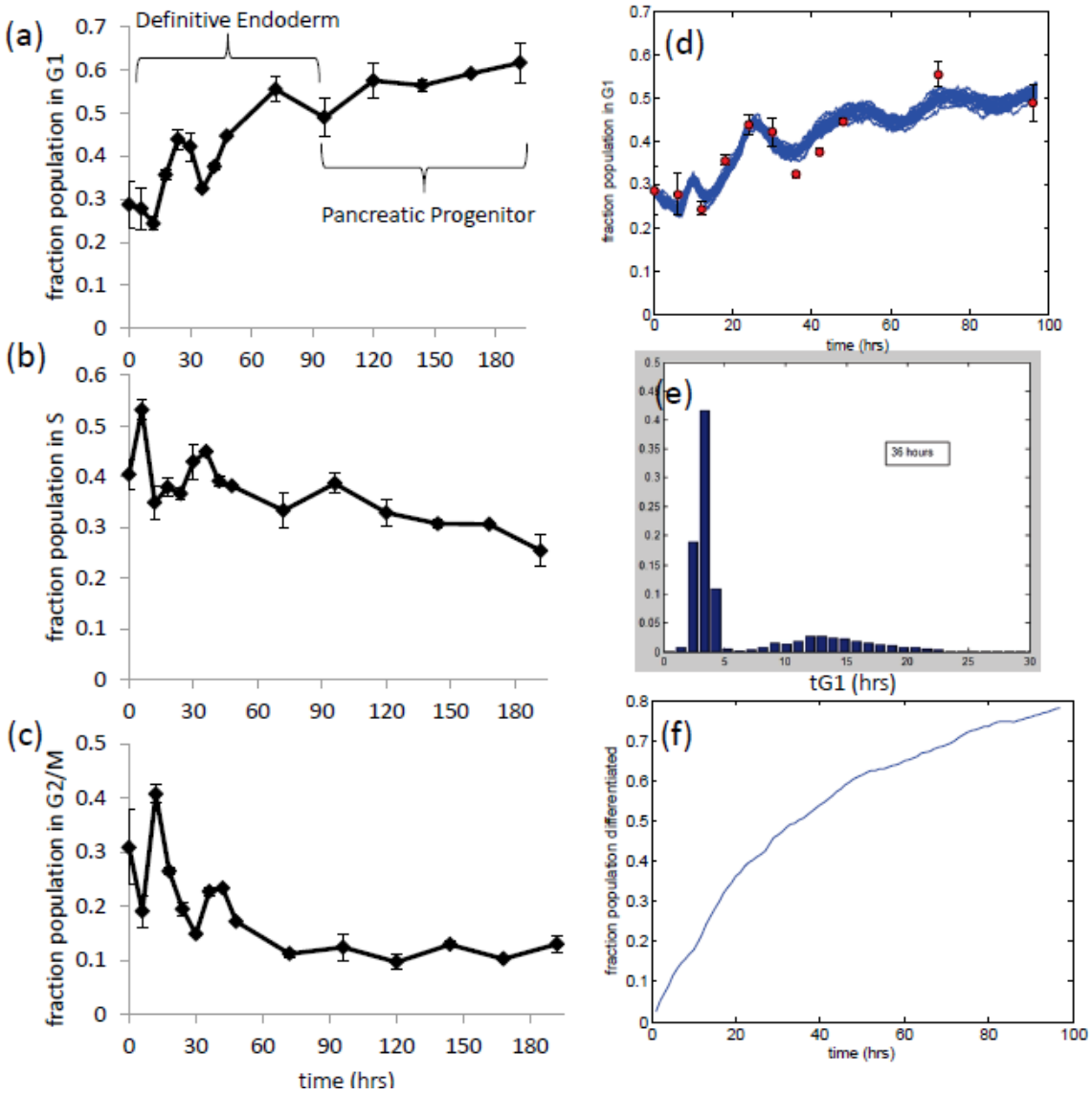

Figure 4.11. Dynamic changes of the H1 hESC cell cycle during pancreatic differentiation

(a-c) experimental analysis of cell cycle distribution during induction towards endoderm (up to day 4) and to pancreatic progenitor (up to day 8). n=2. Error bars=std. dev. (d) Ensemble model predicted dynamics describing G1 changes during endoderm (blue band, 20 stochastic simulations) as compared to experimental data. (e) Snapshot at t=36 hrs of the predicted G1 residence time distribution during the definitive endoderm simulation. (f) Simulated fraction of differentiated cells with time

The ensemble model was applied to the G1 phase definitive endoderm dynamics (Figure 4.11(a), up to 96 hrs). As mentioned before, the model rules were kept largely invariant while the model parameters were re-optimized for the current experimental system. The only difference in the model rule being elimination of tstop parameter when considering DE differentiation. This is needed since the differentiation agents were replenished daily, hence removing the concern of diminished efficacy of the differentiation agent. The model does an excellent job in describing the experimental behavior (Figure 4.11(d)), and predicts relevant cellular parameters (Table 4.3). It is interesting to note that oscillatory behavior was observed in the experimental G1 population dynamics. While the source of these oscillations is non-intuitive, the model is able to capture this behavior. Analysis of the model structure and optimal parameter set revealed the following observation: this oscillatory behavior is likely due to the large jump in G1 length (large γ parameter), and can be explained in the following way. A large "jump" in G1 length, and therefore G1 population fraction, can be thought of "partially synchronizing" the cells in G1. Therefore, as shown in the preceding sections, oscillations will ensue. These will dampen out, but because of the increased G1 time, the new steady state fraction will be higher.

Table 4.3. Parameters associated with the best fit cellular ensemble model to definitive endoderm dynamics

Parameter (Equation 4.8)	Value
α	12.8 hrs
μ_γ	0.305
σ_γ	0.041
Probability of differentiation	28.2%

As with the DMSO system, we analyzed the dynamic changes of the G1 residence time distribution during endoderm differentiation (Figures 4.9(d) and 4.11(e)). In contrast to the DMSO system, for definitive endoderm differentiation, the rate at which G1 phase lengthens is much greater, leading not to a gradual right-tail increase, but an abrupt jump in cells' G1 time (Figure 4.11(e)). This leads to a separate subpopulation with a higher G1 residence time, which may explain the two subpopulations observed in the pancreatic progenitor synchronization model. Further dynamic information can be obtained by analyzing the time dependence of the fraction differentiated cells (Figure 4.11(f)). The simulation shows that there is a steady increase in the proportion of the population which is differentiated, until at around 1 day of directed differentiation the rate at which this proportion increases starts to diminish. By the end of the simulation (Day 4), this fraction seems to reach close to a steady state, with the total percentage of cells differentiated reaching around 80%. This is in agreement with separate flow cytometry analysis, which shows a similar percentage of sox17 positive cells (endoderm marker) at the end of definitive endoderm induction (Figure 4.3(a)). Thus analysis of the cell cycle dynamics using the developed ensemble model offers a powerful tool to characterize differentiation and achieve mechanistic explanations on observed cell cycle behavior, as demonstrated during definitive endoderm commitment.

4.3.5 Changes in single cell protein network account for cell cycle population dynamics and increased variability with differentiation

The developed population model could accurately capture cell cycle dynamics after synchronization, as well as predict phase times and variability, for a given phenotype. We further developed this into an ensemble model, which allowed distinct hypothesis testing of mechanisms

involved in G1 lengthening during the process of differentiation. In the next step we will further advance the ensemble model by accounting for the cell cycle proteins which govern G1-S transition. Numerous proteins interacting in a complex nonlinear network have been demonstrated to govern the cell cycle dynamics and phase transitions. There have been attempts to represent these interactions by a set of coupled nonlinear ODE. [44, 46, 48, 50], primarily targeting eukaryotic system. However, current attempts to elucidate the cell-cycle behavior of ESCs have been exclusively experimental. While a similar model can be theoretically extended to ESC, we developed a molecular model which captures the dynamics of key G1 proteins known to significantly change during differentiation, especially during definitive endoderm commitment. When adapting the base equations describing G1-S protein dynamics ([46, 48]) for use in our system, two considerations were made: (1) a switch in the cyclin E expression is present, which we use as a simplified indicator of G1 transition time; and (2) molecules which have been shown to change during definitive endoderm differentiation (cyclin D [138]) and during differentiation in general (CIP/KIP (e.g. p21, p27) [154-156]), are present. This model is shown in Figure 4.12, with ODE and parameters described in Equations 4.1-4.5 and Table A3, respectively. While this is by no means a full model of the G1-S control circuit, this minimal model provides a mean to capture G1 residence time at the molecular level. The development of this model from a standard G1-S model is outlined in section 4.2.5. Profiles of the three proteins at nominal (undifferentiated) values are shown in Figure 4.12 (b).

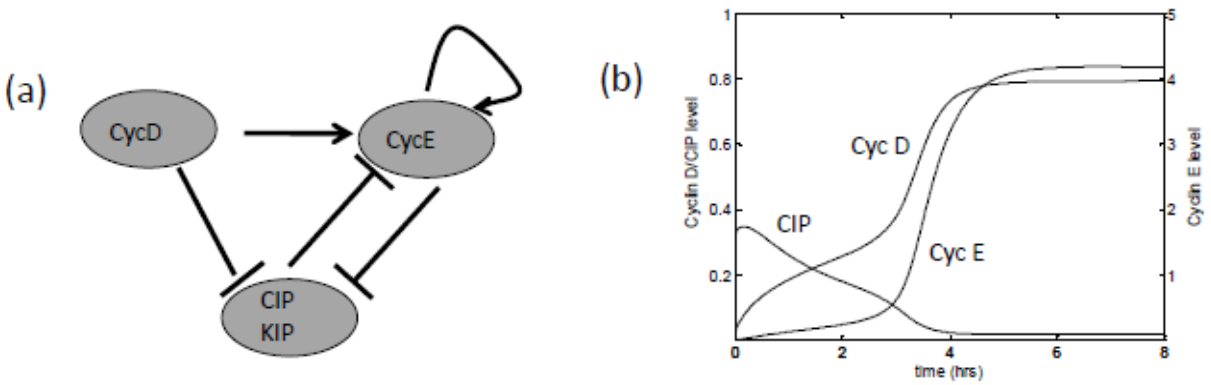

Figure 4.12. Single cell ODE model

(a) Schematic of the simplified protein network governing G1 residence time. (b) Protein temporal profiles simulated from the single cell ODE model under nominal (undifferentiated) parameters.

Two different systems were tested: DMSO differentiation by Calder *et al.* [41] and definitive endoderm differentiation. For the former, it has been shown that a mechanism by which DMSO may lengthen the G1 phase is by inhibiting phosphorylation of the Rb protein [153]. This would suggest that the efficacy of cyclin D to phosphorylate Rb, thereby releasing E2F with subsequent promotion of cyclin E, would be diminished (for the logic of these protein connections, see [46, 48]) (Figure 4.13(a)). While Rb and E2F are not captured in Equations 4.1-4.5, we can simulate this effect by increasing the Michaelis-Menten parameter associated with the D→E term (parameter k9 in Equation 4.2, Figure 4.13(b)). To simulate DMSO effect in our simulation, the k9 parameter was increased linearly with time (Figure 4.13(c)). This effect, at the molecular level, is shown in Figure 4.13(d-f). As shown, the time at which cyclin E jumps from a low level to a high level, represented in this model G1 transition time, increases with increasing parameter value, thereby demonstrating how DMSO increases G1 residence time at the single cell level. Each profile represents a different value of k9, with the value increasing linearly. It is interesting to note that this linear increase in k9 results in an exponential-like increase in G1

transition time (Figure 4.13 (g)). This is in agreement with the mechanism proposed by the generalized model (Equation 4.8), which predicts that G1 time increases exponentially with time. These ODE were incorporated into the population model, and replaced the exponential G1 lengthening in Equation 4.8. To account for cellular population heterogeneity, the amount k9 increased is not deterministic, but has a stochastic component. Therefore, the parameters which were estimated were the same as in the aforementioned generalized model (Equation 4.8), including tstop, maximum G1, probability of differentiation, and the mean and variability associated with G1 increase. However, with the latter two parameters, rather than γ, in the current ODE formulism we are estimating the mean and variability of k9. The resulting population dynamics, and agreement with experimental data, are shown in Figure 4.14 (a-c).

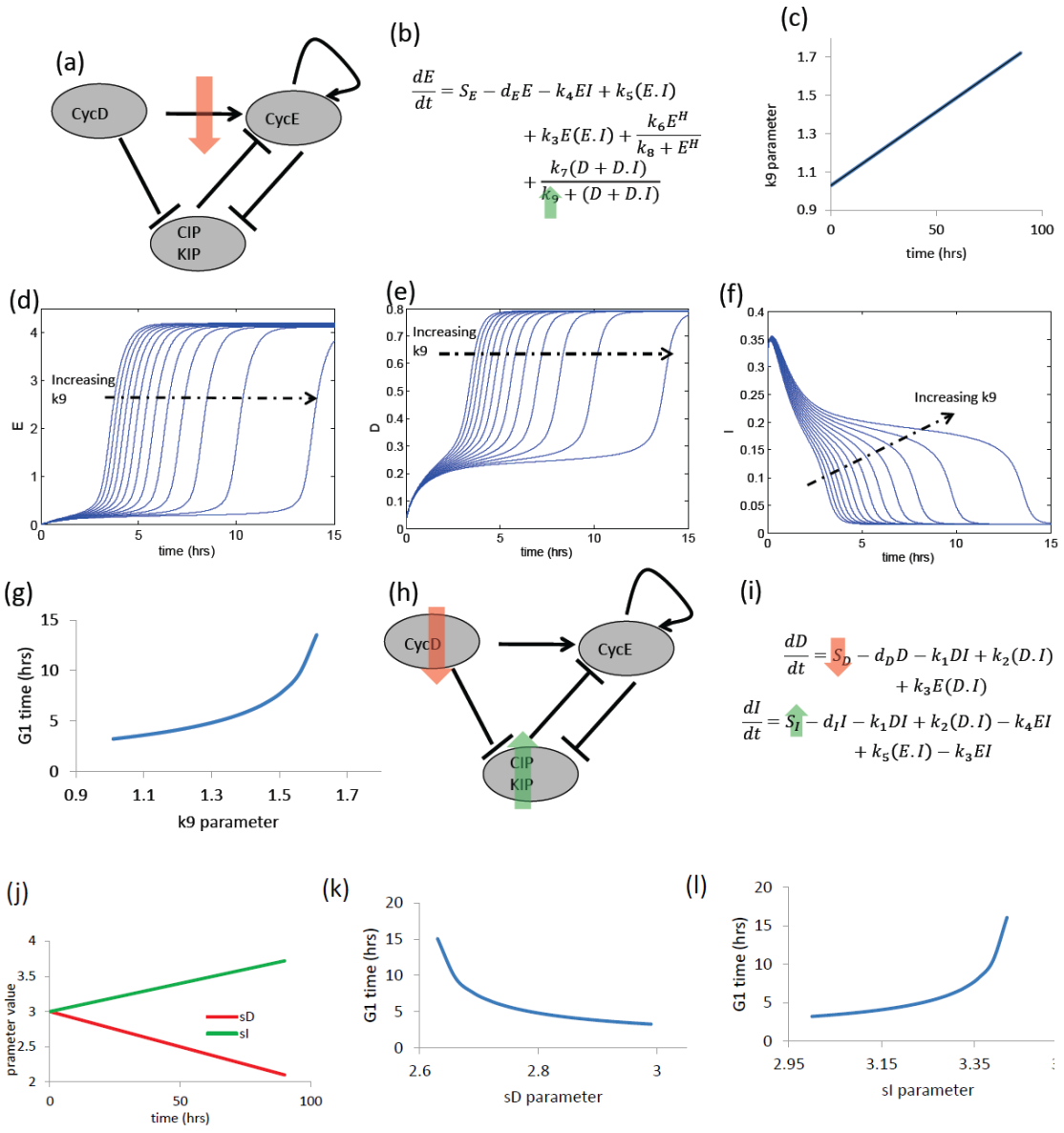

Figure 4.13. Effects of different perturbations on the G1 protein behavior

(a) Representation of the effect of DMSO on the protein network, which is captured by increasing parameter k9 in Equation 4.2 (b) linearly with time (c). (d-f) profiles of cyclin E, cyclin D, and CIP/KIP, respectively, with a linear increase in parameter k9. (g) G1 residence time as a function of k9 parameter. (h) Representation of the effect of definitive endoderm differentiation on the protein network, which is captured by decreasing parameter s_D in Equation 4.1 and increasing parameter s_I in Equation 4.3 (i) linearly with time (j). G1 residence time as a function of s_D parameter (s_I at nominal value) (k) and s_I parameter (s_D at nominal value) (l)

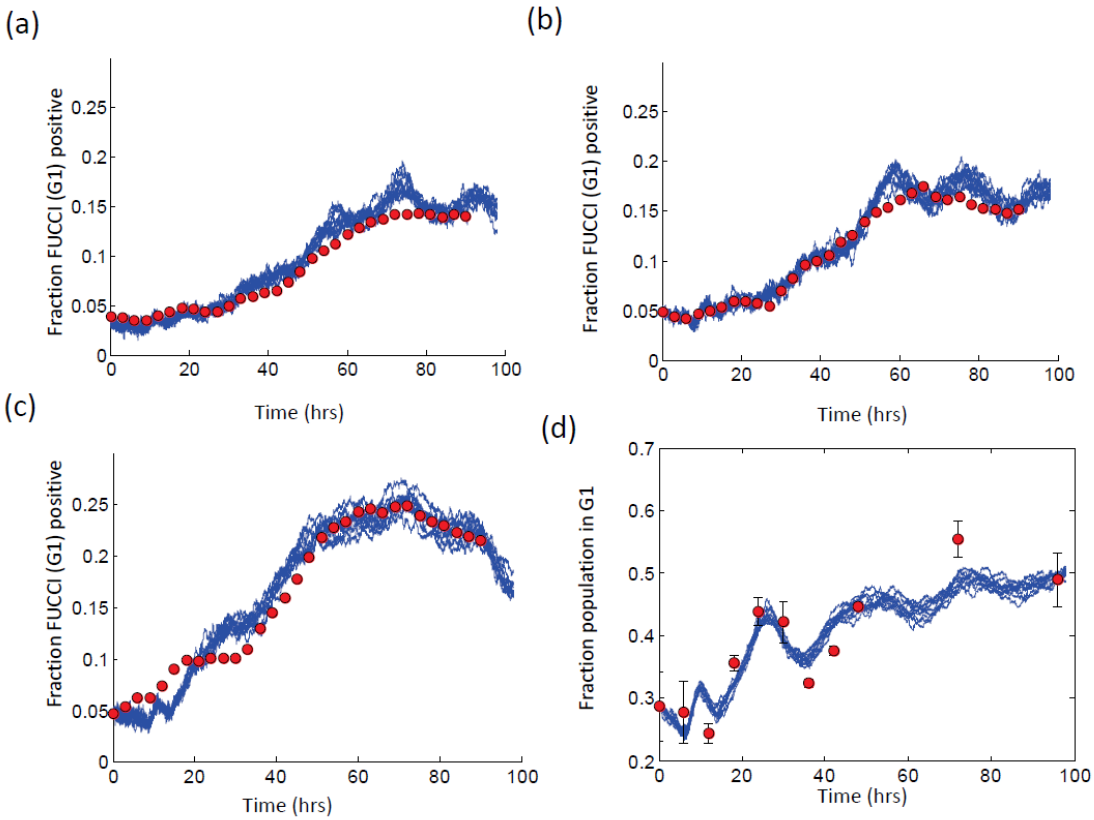

Figure 4.14. Application of ODE ensemble model to G1 population dynamics induced by various induction conditions

(a-c) 0.5%, 0.75%, and 1% DMSO, respectively; (d) definitive endoderm. (error bar = st. dev.). Shown are 10 stochastic runs

This model was also applied to definitive endoderm differentiation. Numerous studies have reported that during differentiation, the levels of G1 inhibitors rise [154-156], and more specifically, during endoderm commitment, the level of cyclin D goes down [138]. Figure 4.13(h) displays how this was simulated in our ensemble model, including a decrease in Cyclin D and an increase in CIP/KIP. In Equations 4.1 and 4.3, this can be achieved by decreasing and increasing sD and sI parameters, respectively (basal synthesis of cyclin D and inhibitor, (Figure 4.13(i)). These synthesis rates changed linearly with time (Figure 4.13(j)). As above, a linear

change in these molecules gives rise to an exponential increase in G1 time (Figure 4.13(k,l)). When incorporated into the population model, the linear changes of these parameters included a random component, again to account for population heterogeneity. The simulated model dynamics and agreement with experimental data are shown in Figure 4.14(d) and displays the efficacy of the model to capture definitive endoderm cell cycle population dynamics resulting from perturbations at the single cell level.

Given the established model, one can perform hypothesis testing and make predictions on how different perturbations may affect cell cycle behavior. One important aspect of the stem cell system is cell-to-cell variability, and how heterogeneous populations arise. Our ODE model provides us with a platform to analyze heterogeneity. Specifically, we wanted to determine how intracellular variability, including fluctuations in protein levels, contributes to cell cycle population variability with respect to cell cycle, and if this changes upon differentiation. To accomplish this, we first simulated the self-renewing system with nominal parameters (Table A3) and various coefficients of variation (CV) for both cyclin D and CIP/KIP synthesis levels (sD and sI, respectively), and quantified the distribution of G1 residence times. We then performed the same task with a differentiated cell type by adjusting these synthesis levels to yield a G1 residence time of 12 hours (either μ_{sD}=2.642 mM/hr or μ_{sI}=3.402 mM/hr). Figure 4.15 shows the results of this analysis. While the trends are similar between cyclin D and CIP/KIP, the same cannot be said regarding the comparison between different stages of maturation. With undifferentiated parameters, the system behaves linearly, with a 10-fold increase in the parameter CV giving rise to a 10-fold increase in the G1 time CV. Interestingly, this linear correspondence is not observed with the more differentiated system; the 10-fold increase in parameter CV gives rise to a 30-fold increase in the CV of G1 times. Furthermore, given the same parameter

variability, the corresponding G1 time variability is much higher in the more mature system. Another striking feature is the shape of the distribution itself. With an undifferentiated cell type, and with very low parameter variation with a differentiated cell type, the G1 residence time distribution seems to be normally distributed. However, with higher amounts of variability, this distribution becomes more skewed, resembling a log-normal or gamma distribution, which is in agreement with the generalized ensemble model (Figure(4.9 (a-c)). The results suggest that heterogeneity associated with differentiation could in part be due to the increased variability of G1 times with maturation, and demonstrate the predictive utility of the model.

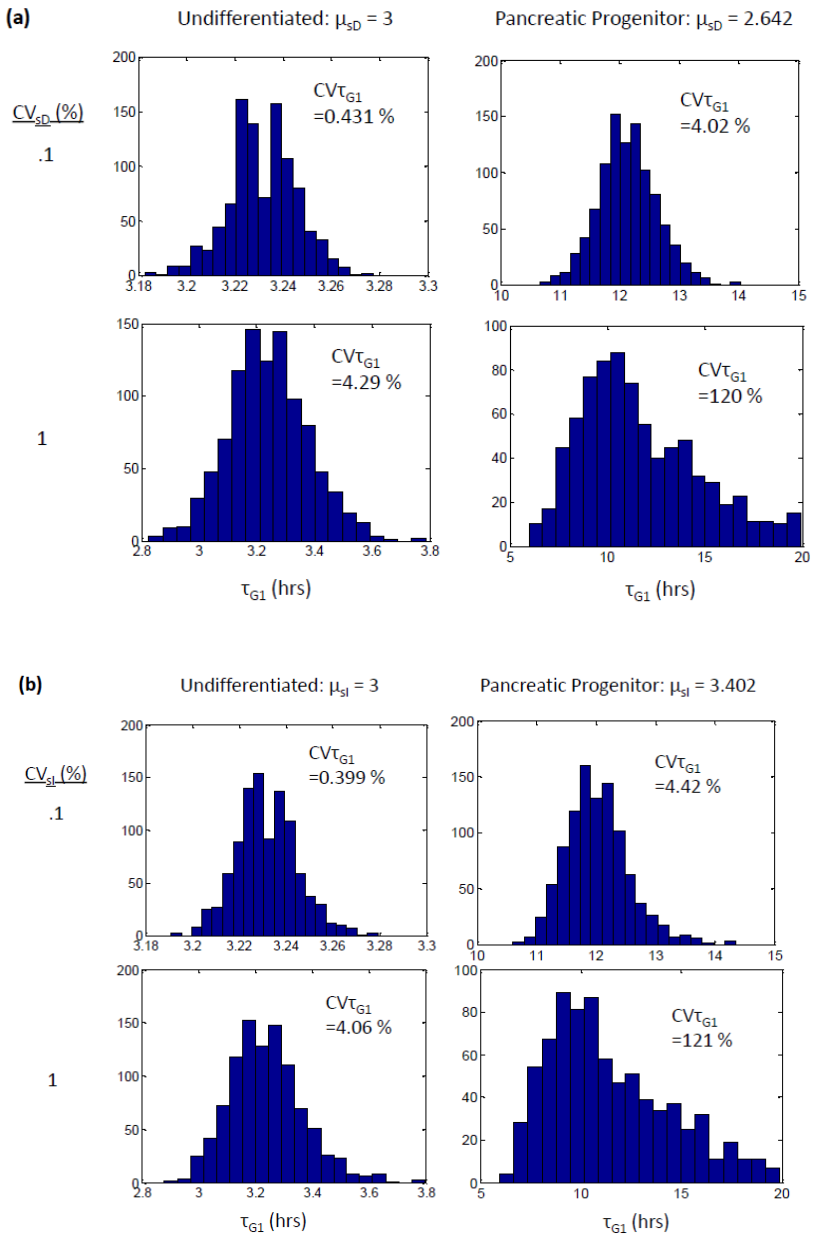

Figure 4.15. Effect of ODE parameter variability on G1 residence time variability

Synthesis rates of cyclin D (s_D) (a) or CIP/KIP (s_I) (b) were normally distributed with different coefficients of variation (CV) (The variability of the parameter not being distributed (e.g. s_I in (a) and s_D in (b)) was 0). For each value, Equations 4.1-4.5 were solved and G1 times determined. This was performed with mean synthesis values yielding a relatively short mean G1 time (~3.2 hrs, undifferentiated hESC, left panels) and a longer mean G1 time (~12 hrs, right panels)

4.4 DISCUSSION

Mathematical models of the cell cycle have been of great interest in order to understand the dynamics and mechanisms behind cell growth and proliferation, and have been applied to systems ranging from yeast to cancer cells. The cell cycle of the embryonic stem cell system has recently gained much attention, and over the past few years there has been a concerted experimental effort to understand the state of the cell cycle proteins contributing to the unique behavior of mouse and human ESC. However, a quantitative understanding of the cell cycle in ESC is still lacking. This work is the first attempt to mathematically represent the dynamics of human ESC cell cycle behavior and achieve a more thorough quantitative understanding of the system.

In this work we have utilized a coupled experimental and computational approach to quantitatively represent the unique cell cycle behavior of hESC during self-renewal and differentiation. Cell cycle dynamics were first analyzed through synchronization experiments which show drastic differences between cell cycle behavior of self-renewing hESC and hESC-derived pancreatic progenitors. These dynamics were captured by a stochastic population model, allowing for prediction of relevant cell cycle characteristics, including population heterogeneity and the phase residence time distributions. To capture the process of and elucidate the mechanisms behind G1 population increase during the process of differentiation, we developed an ensemble model which accounts for differentiation and G1 lengthening at the single cell level. This model was then extended to account for intracellular protein dynamics, giving insight into the mechanisms of G1 lengthening and population variability.

Numerous approaches have been developed to mathematically understand cell cycle population dynamics, with a primary focus being analysis of cancer cells and the effects of drug

therapies. These approaches range from relatively simple ordinary differentiation equations to complex spatially and physiologically structured partial differential equations [157]. Herein we have adopted an approach which describes phase population and synchronization/desynchronization behavior by tracking cellular automata, whose phase times are described by probability distributions. There were several reasons for adopting this approach. The first is the directness by which stochasticity and variability can be incorporated, which was a focal point of the analysis, and which has been shown to be crucial in cell cycle analysis [141, 158, 159]. Also, others have demonstrated the effectiveness of this modeling technique for mechanism prediction [54, 55], which was another goal of the current work. A more common modeling technique to describe population dynamics is the population balance equation (PBE), which has been used to model various systems, ranging from adult stem cell behavior [35] to plant cell aggregation [160]. PBE have also been used to study cell cycle dynamics, including models whose phases have age dependence [61], analogous to the phase residence times in the current work. However, these models often result in integro-partial differential equations, and while there have been advances in approaches to solve these equations, including discretization techniques, if the number of variables describing the intracellular state becomes more than one or two, it may not be possible to realistically solve them at all [36]. As one of the objectives of this paper was to describe the state of the cell by intracellular cell cycle proteins and the heterogeneity associated with these parameters, this approach would not be feasible, and we therefore adopted the single-cell approach, which can easily incorporate many variables to describe the cellular state.

Recent years have seen advances in single cell analysis, which can yield more information of a cellular population than population-averaged data [161-163]. This is true for the

cell cycle system; FUCCI reporters, flow cytometry, single cell time-lapse tracking and mitosis detection are able to provide a wealth of information, including quantification of heterogeneity and individual G1/cell cycle times [41, 143, 144]. While it is possible to experimentally track single cell and population dynamics with ever-increasing resolution, this data alone is often inadequate in developing a mechanistic understanding of the complex system. Hence, we coupled single-cell information to our developed population and ensemble models to achieve this mechanistic understanding. A strength of this model is the ability to perform hypothesis testing, something which is not possible with experimentation alone. Through hypothesis testing, we are able to both elucidate mechanisms of cell cycle transition and make informed predictions of the effects of different perturbations to cell cycle behavior.

Several researchers have suggested that G1 times are exponentially distributed [141, 151]. We therefore expected our undifferentiated model and synchronization experiments to predict this behavior. This was not the case, with the best agreement arising from purely Gaussian distributions. This was not expected, and we hypothesize that at this pluripotent state the control mechanisms in place are such that there is tight control on the G1 phase which prohibits cells from probabilistically exiting the cell cycle and obtaining longer times, thereby maintaining fast proliferation and minimal time in which cells are exposed to differentiation cues. We postulate that as cells mature, this tight control becomes more relaxed, allowing for longer and more variable G1 times, which might be necessary with a more mature system. This transition would cause the normal distribution to become positively skewed, resembling a more gamma or exponential distribution. Indeed, this can be seen happening from the ensemble model simulating DMSO induced differentiation (Figure 4.9(a-c)).

A further unexpected result came when analyzing the pancreatic progenitor system. First, as one can see from Figure 4.3(c), a high degree of synchronization was not achieved in this system. This was surprising, given the high degree of synchronization at the self-renewal stage (Figure 4.2(c)). Even with this less than optimal degree of synchronization, the flexibility of the model allows for the cell cycle dynamics to be captured and relevant parameters extracted, as long as the cell cycle distribution has been at least somewhat perturbed from its asynchronous state. It should be noted, however, that more accurate model predictions may be obtained with increasing synchronization. In addition to experimental synchrony, another important observation in the pancreatic progenitor system was the emergence of two populations of cells. When considering two, as suggested by DFT, excellent agreement between the simulation and data was obtained. This was corroborated through the ensemble model which predicted a population of cells displaying a different set of G1 traits, which arose from the predicted large G1 increase during endoderm differentiation. Furthermore, the percentage of differentiated cells at the end of the endoderm simulation (Figure 4.11(f)) is in relatively close agreement to the percentage of the population displaying higher doubling times predicted by the pancreatic progenitor synchronization model (61%, Table 4.2(b)). While the ensemble model was trained on a completely separate data set, and was not given any information on multimodality of G1 times in the population, this multimodal distribution was nevertheless predicted (Figure 4.11(e)), further demonstrating the effectiveness of the modeling techniques. While this ensemble simulation was not continued throughout the pancreatic progenitor differentiation process, it is thought that these two modes would persists throughout, albeit shifted through further changes brought about by pancreatic differentiation.

In analyzing the phase residence times, and the changing of these times during differentiation, we particularly focused on the G1 phase. While there may be changes in the S and G2/M phases, at the current time the consensus in the field is that the majority of the change to the cell cycle during differentiation happens in the G1 phase [143]. Furthermore, it is a widely held belief that the majority of cell fate decisions, especially those involving ESC behavior, are made in the G1 phase [137, 164]. We therefore focused on this phase by which population dynamics is mostly affected. In the next step, when building our reduced G1 molecular ODE model for the ensemble model, we focused on those molecules in early G1, and those which are present in late G1 (guiding the cells towards DNA replication) were not considered. Studies have shown that the early part of G1 is subject to high variability, with the latter part more deterministic [141, 151]. Therefore, we hypothesized that changes in cell cycle length and variability are associated with the early G1 phase. In addition, it has been postulated that differentiation, especially towards endoderm, is more likely to happen in early G1 [138, 142]. Early G1 control was therefore chosen to be a focal point of the ODE model and for G1 lengthening.

During the development of the ensemble model, parameter estimation, and elucidating mechanistic behavior, it is important to note that not all model alternatives worked, and the optimal model described herein was the only one to capture the unique dynamics of the G1 population upon differentiation. A crucial aspect of the physical behavior that allows for this process to work is the nonlinearities associated with the temporal G1 population data, including the delay, step behavior, small and large oscillations, and end of experiment decay. If these features were not present, and the differentiation data resulted in simple monotonic, linear behavior, there would most assuredly be redundancy, with multiple models being able to explain

the experimental dynamics. However, with these nonlinearities, only one model alternative was able to do so, with the elucidation of mechanistic information. One of these mechanistic determinations was the necessity incorporating G1 “priming” to describe the population dynamics. Cells decide to differentiate only in the G1 phase, and if the cells do commit to differentiate, lengthening occurs upon entering G1 phase of the next cell cycle. This theory of G1 “decision making” has been postulated through experimental evidence of several groups [41, 137, 138, 142], as has the idea that G1 lengthening does not occurs immediately [150]. This behavior was suggested by comparing model alternatives to population data, without any direct experimental data on G1 decision making, further demonstrating the power of the modeling approach.

When analyzing the cell cycle of stem cells, it is important to carefully separate the characteristics associated with adult and embryonic. While there has been considerable effort in studying, both experimentally and computationally, the adult stem cell cycle [165-168], the features associated with this behavior may or may not be directly translatable to studies looking at ESC division. For instance, asymmetric division and quiescence, hallmarks of adult stem cells, have not been characterized in ESC. In looking at this latter point in detail: many works modeling the G1-S switch at the molecular level concentrate on quiescence, in particular a G0 phase in which the cell exits the cell cycle due to a lack of cues to move forward in the cycle, namely growth factors. However, it is currently not known if this restriction point exists in the ESC cell cycle, and if checkpoints are present, research is ongoing to determine in what fashion they operate [169-171]. Furthermore, it has been shown that ESC exposed to low serum conditions do not exit the cell cycle, and continue to proliferate [150]. Therefore, the idea of a

G0 phase may not be applicable to ESC cell cycle studies, and a more gradual lengthening of the G1 phase, induced by different factors, as modeled in the current work, may be more applicable.

In addition to differences between adult and embryonic, differences between species should be taken into account. Genes and pathways are often conserved through evolution; indeed, studies have shown that the control mechanisms governing the G1-S transition have similarities between eukaryotes [172]. However, there seem to be substantial differences between the cell cycle of mESC and hESC. First, the doubling time in mESC is even shorter in mESC, ~8-10 hrs, with a smaller proportion of the cell cycle in the G1 phase, ~15% [173]. This could be due to another substantial difference: elevated CDK activity which is not dependent on cell cycle position [174, 175]. If this were the case with hESC, the ODE model proposed herein would not be applicable. This, however, does not seem to be the case, as others have reported that, depending on the state of the cell cycle, proteins governing cell cycle transitions can be expressed at different levels and with different activities [154, 176, 177].

4.5 CONCLUSIONS

In this study we have developed a stochastic population model to describe and understand the cell cycle system of hESC during self-renewal and differentiation. We started off with a stochastic cellular automaton model; this tracks individual cells and their progress through the cell cycle, with phase residence times described by probability distributions. Information on the system is obtained by comparing simulations to experimentally synchronized cells, and accurate predictions of phase resident time distributions of undifferentiated cells are achieved, including phases which are governed by tight Gaussian distributions. This analysis was extended to explain

the desynchronization of pancreatic progenitor cells. The predicted phase distributions show that, in this more mature phenotype, their exists two separate subpopulations which display different cell cycle behavior, but with both populations showing G1 phases longer and more variable than undifferentiated cells. To mechanistically explain this behavior, we developed an ensemble model of G1-S transition, describing how this G1 lengthens upon differentiation. This is achieved in two parallel ways, with one model consisting of a phenomenological formulation, and the second as a set of ODE describing the temporal profiles of proteins controlling the cell cycle. Both of these structures were incorporated into the cellular automata. This ensemble model accurately captured the change in G1 population during differentiation and predicts pertinent features of this transition, including probability of lengthening, maximum G1 time, and the mechanism by which multiple cell cycle populations arise.

5.0 POPULATION BEHAVIOR OF STOCHASTIC CELLULAR DECISION MAKING DURING INITIAL LINEAGE COMMITMENT

5.1 INTRODUCTION

Chapter 4 demonstrated that mechanisms governing G1 lengthening at the population level can be determined through our developed stochastic model. To gain a more mechanistic understanding on the differentiation process itself, we extend this model to focus on population dynamics arising from initial lineage commitment to the germ layers [178]. We show that through this model, a better understanding of the endoderm differentiation process of human ESC can be obtained.

Overall, the extended model involves a stochastic simulation, where a population of cells is evolved following specific rules through which the system dynamics can be extracted. The model predicts three aspects of endoderm formation: total cell proliferation, cell death, and lineage commitment. In order to understand the mechanism favoring the process of stem cell differentiation we simulate several alternate mechanisms and compare the simulated dynamics with our experimental data. Endoderm is experimentally induced in hESC through alternate pathways: addition of Activin A and Activin A supplemented with the growth factors basic fibroblast growth factor (FGF2) and bone morphogenetic protein 4 (BMP4) [15, 179, 180]. Differentiation dynamics of the cell population is experimentally tracked by analyzing

percentage of cell population expressing endoderm specific proteins: Sox17 and CXCR4 [181, 182]. Through agreement between the experimental data and simulated dynamics, we elucidate the mechanisms behind initial lineage commitment during definitive endoderm differentiation induction.

5.2 METHODS

5.2.1 Cell culture and endoderm induction

The culture of H1 hESC was described in Chapter 4. For the differentiation study, the current work chose to compare two conditions to induce endoderm: human Activin A (henceforth referred to as 'Condition A') and human Activin A supplemented with the growth factors FGF2 and BMP4 ('Condition B'). To commence endoderm induction, DMEM/F12, 1xB27® Supplement, and 0.2% Bovine Serum Albumin (BSA) supplemented with 100 ng/mL human Activin A (Condition A) or 100 ng/mL human Activin A, 100 ng/mL FGF2, and 100 ng/mL BMP4 (Condition B) was used as differentiation media, which was replaced daily for a total of five days. Upon induction of differentiation cells were harvested on a daily basis for subsequent analysis. For each well, the supernatant was collected, the plated cells were dissociated with Trypsin+EDTA, Trypan Blue was added to distinguish live from dead , and the cells were counted using a standard hemacytometer.

5.2.2 Flow cytometry and quantitative polymerase chain reaction

For flow cytometry, harvested cells were first fixed for 15 minutes in 4% methanol-free formaldehyde in phosphate-buffered saline (PBS). Cells were washed twice and permeabilized in 0.1% Saponin + 0.5% BSA in PBS for 30 minutes. To block non-specific binding, the cells were incubated in 3% BSA + 0.25% dimethyl sulfoxide (DMSO) + 0.1% Saponin in PBS for 30 minutes. A portion of cells were then set aside as the negative control (secondary antibody only without primary). The cells to be used as the positive samples were then incubated in blocking buffer with goat anti-human sox17 and rabbit anti-human cxcr4 primary antibodies, 1:200 dilution, for 30 minutes. The cells were washed twice with blocking buffer, resuspended in the buffer, and incubated with donkey anti-goat APC (1:350 dilution) and donkey anti-rabbit FITC (1:200 dilution) for 30 minutes (both the samples and negative control). Two washings were followed by 10 minutes of 0.2% tween-20 to further eliminate non-specific staining. Cells were washed and transferred to flow cytometry tubes. Accuri C6 © Flow Cytometer was used to quantify sox17-APC and cxcr4-FITC expression. Cells stained with the secondary antibody only (without primary antibody) were first analyzed; this population was taken as the negative, and the gate was set beyond these cells to eliminate false positives due to auto-fluorescence and non-specific secondary antibody binding. The completely stained samples (primary and secondary antibody stained) were then analyzed, and the percentage of the population falling within the set gate was recorded as the positive sample for the respective antibody.

To quantify mRNA levels, harvested cells were lysed and mRNA was extracted and purified using a Nucleospin II RNA kit. The RNA quantity and quality was measured using a SmartSpec^{TM} Plus spectrophotometer , after which reverse transcription was performed with the ImProm II Reverse Transcriptase System. cDNA levels of Gapdh, Oct4 , and Brachyury were

measured with quantitative polymerase chain reaction (qPCR) using an Mx3005P system and Brilliant SYBR Green qPCR Master Mix (primers used in qPCR shown in Table A5).

5.2.3 Mathematical model

The system of stem cell differentiation to endoderm is modeled using a stochastic population-based model. The basic formulation of the model is based on earlier reports for hematopoietic stem cells [37, 183]. Here we are introducing some modifications to adapt it to the ESC system, followed by a stringent model analysis using parameter sensitivity and feasibility studies. In this section we summarize the working principle of the model along with our modifications. Figure 5.1 displays the pseudo-code describing the implementation of the main simulation, parameter ensemble, and sensitivity analysis.

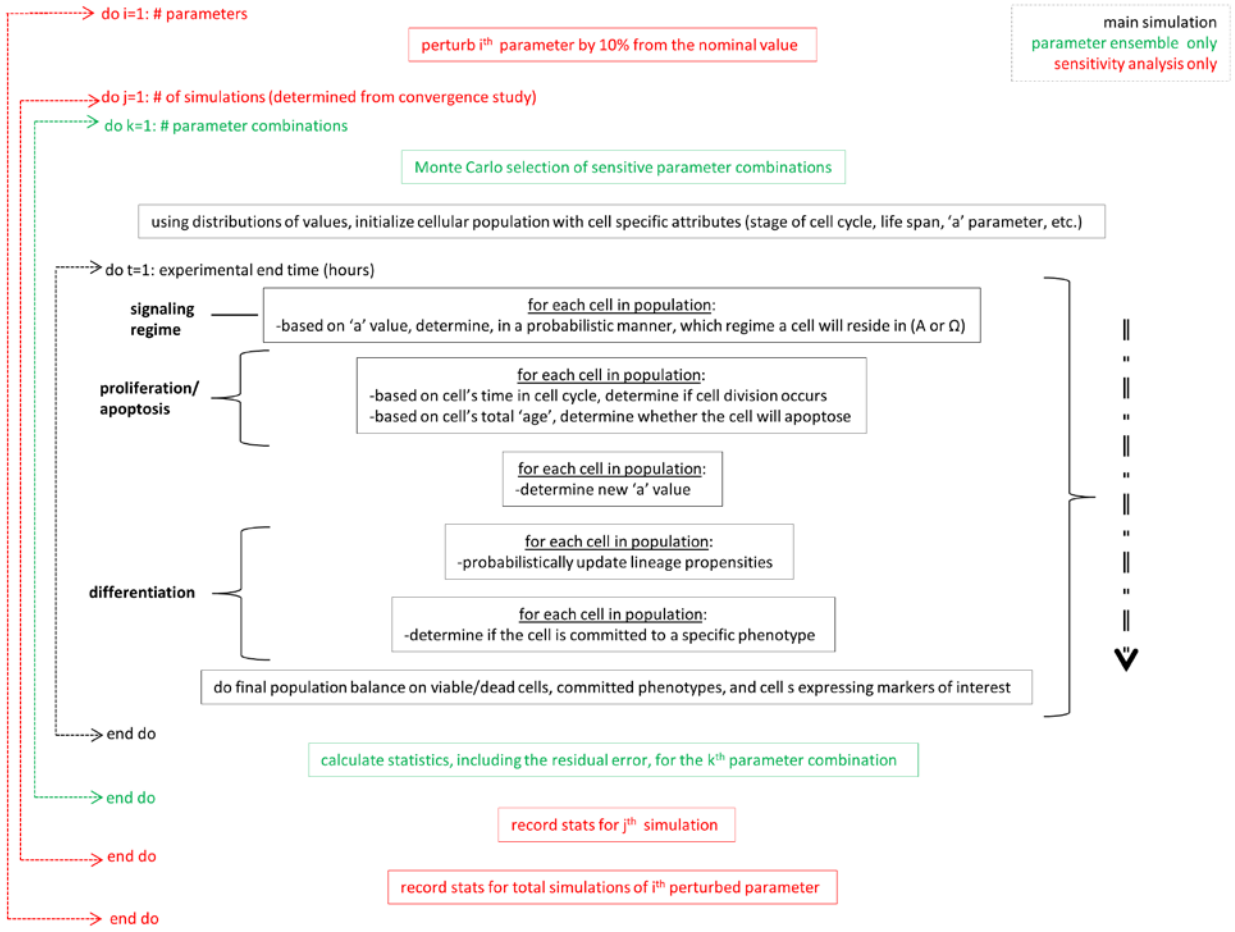

Figure 5.1. Implementation of mathematical model

Pseudo-code describing flow of events in the population based model. Black: events to simulate temporal behavior of cellular population (main routine). Green: model inclusions for parameter ensemble, which runs main routine using different parameter value combinations. Red: model inclusions for sensitivity analysis, which runs main routine 4000 times (replications determined by convergence study) for each perturbed parameter value, the output being the parameter sensitivities.

5.2.3.1 Signaling Regimes, proliferation, apoptosis, differentiation rules

The model is initiated by a population of cells, the properties of which are evolved by specific pre-assigned rules. The cells are primarily categorized into two signaling regimes: Ω and A; Ω can be considered as an active regime supporting cellular proliferation and differentiation, while A is a more dormant regime where cells are prone to dedifferentiation. The cells can transfer in between these two regimes, an event decided upon primarily by a cell-specific parameter, termed as 'a' value. This parameter 'a' is randomly assigned to each cell at the beginning of simulation, and is updated at each time step. The probability of transfer to/ from a regime is dependent on this parameter along with the number of cells in the destination regime. This 'a' value is unaltered in the A regime while it progressively reduces in the Ω regime, and when it falls below a threshold the cells lose their ability to transfer to the A regime.

Each cell is randomly assigned a maximum life span, exceeding which it will die. While the cells age in the Ω regime, they neither proliferate nor age in the A regime. Proliferation is allowed in the Ω regime for an amount of time which is cell dependent, after which the cell enters a senescent stage and will not proliferate. Furthermore, an individual cell is allowed to proliferate only after it loses the capability to pass into A regime having crossed the 'a' threshold value.

Cellular differentiation is governed by the 'lineage propensity' parameter, representing a cell's likelihood to differentiate into a particular lineage. Only the Ω regime allows an increase in lineage propensity. While updating the propensity of differentiation to a particular lineage, all the possible lineages are competing and any can be updated, with the one with higher propensity having a higher probability of being selected. Once a cell's propensity exceeds a threshold level, the cell is considered committed to that particular lineage and will retain its differentiated

phenotype. If this threshold has not been exceeded and if the cell is chosen to be transferred to the A regime, the propensity values will converge to an average value. The model is therefore able to track specific germ layer populations, and through this the percent of the population positive for Sox17 (visceral and definitive endoderm marker) and CXCR4 (definitive endoderm and mesendoderm marker) [181, 182, 184] can be extracted.

5.2.3.2 Mechanism of hESC differentiation

The current work is focused on the mechanistic investigation of the dynamics of hESC induction into endoderm. Using the platform of the stochastic population based model we investigated several alternate mechanisms and analyzed them for agreement with experimental data. Three characteristics of the differentiation process were chosen to be investigated: the presence/absence of an intermediate germ layer, mesendoderm, which subsequently gives rise to mesoderm and endoderm [185]; the presence/absence of CXCR4, in mesoderm; and whether proliferation of a specific differentiated cell phenotype is favored over others. Combination of aforementioned attributes results in 12 alternate mechanisms, shown below:

1. All phenotypes with increased proliferation, CXCX4 in mesoderm, mesendoderm intermediate germ layer present
2. Endoderm and undifferentiated phenotypes with increased proliferation, CXCX4 in mesoderm, mesendoderm intermediate germ layer present
3. Undifferentiated phenotype with increased proliferation, CXCX4 in mesoderm, mesendoderm intermediate germ layer present
4. All phenotypes with increased proliferation, CXCX4 not in mesoderm, mesendoderm intermediate germ layer present

5. Endoderm and undifferentiated phenotypes with increased proliferation, CXCX4 not in mesoderm, mesendoderm intermediate germ layer present
6. Undifferentiated phenotype with increased proliferation, CXCX4 not in mesoderm, mesendoderm intermediate germ layer present
7. All phenotypes with increased proliferation, CXCX4 in mesoderm, mesendoderm intermediate germ layer not present
8. Endoderm and undifferentiated phenotypes with increased proliferation, CXCX4 in mesoderm, mesendoderm intermediate germ layer not present
9. Undifferentiated phenotype with increased proliferation, CXCX4 in mesoderm, mesendoderm intermediate germ layer not present
10. All phenotypes with increased proliferation, CXCX4 not in mesoderm, mesendoderm intermediate germ layer not present
11. Endoderm and undifferentiated phenotypes with increased proliferation, CXCX4 not in mesoderm, mesendoderm intermediate germ layer not present
12. Undifferentiated phenotype with increased proliferation, CXCX4 not in mesoderm, mesendoderm intermediate germ layer not present

Each of these were incorporated into the model and analyzed for agreement with experimental data. It is expected that the most likely mechanism will best describe the experimental dynamics of the stem cell system.

The incorporation of mesendoderm involved a two stage differentiation scheme. In the first stage, hESC are able to differentiate into either mesendoderm or visceral endoderm. Once cells are committed to the mesendoderm lineage, several of their attributes are re-initialized, such

as their ‘a’ value and lineage propensities. The mesendodermal cells can then further differentiate into endoderm or mesoderm. Differences in the proliferation potential of different phenotypes were incorporated by considering 3 scenarios: proliferation of hESC and mesendoderm; proliferation of hESC, mesendoderm and endoderm; and proliferation of all phenotypes.

5.2.3.3 Convergence study, stochastic sensitivity analysis, and parameter ensemble

As with any stochastic model, the number of model runs necessary to obtain a converged solution needs to be determined. An additional parameter of the current model is the initial cell population, which affects the solution over a certain range. A two-dimensional convergence study was undertaken, wherein the effects of both stochastic run number and initial cell population on model output were determined. The convergence test allows determination of the minimum number of stochastic runs and initial cell population beyond which the model output does not significantly change. All results reported here are using the converged parameter values.

Sensitivity analysis was performed to determine the relative importance of parameters in affecting the outputs of cellular growth, death, and lineage commitment. Because this model is probabilistic in nature, traditional ways of determining local sensitivity, e.g. partial derivative of output with respect to an input, cannot be employed. A stochastic analysis was therefore chosen, using differences in output histograms of nominal and perturbed parameters, S, to estimate parameter sensitivity [186] as:

$$S = HD = \sum_{i=1}^{k} \left| \frac{\sum_{j=1}^{|x|} X(x_j, I_i)}{|x|} - \frac{\sum_{j=1}^{|y|} X(y_j, I_i)}{|y|} \right| \tag{5.1}$$

where x_j and y_j are the individual elements in the nominal and perturbed histograms, respectively, I_i represents the range of each bin i in the nominal histogram, X is a counting variable which takes on a value of 1 if $x_j y_j$ is within the interval I_i, |x| and |y| are cardinalities of data sets x (nominal) and y (perturbed), respectively, and k is the number of bins, which was determined by calculating the appropriate bin size of the nominal output histogram by the Freedman-Daconis rule [187]:

$$\text{Bin size} = 2IQR(P)n^{\frac{-1}{3}} \tag{5.2}$$

Where IQR(P) is the interquartile range of a sample population P and n is the number of observations in P. The number of bins was different for each output, and ranged from 37 to 51.

For sensitivity analysis, each parameter was perturbed by 10% while keeping the rest at the nominal value. For each bin in the nominal output histogram, the difference between the percentage of total nominal histogram elements residing in that bin and the percentage of the total perturbed histogram elements residing in that bin is calculated. The sum of the absolute value of this difference over all of the bins is the histogram distance.

Having determined the most sensitive model parameters, the next step is to determine an appropriate value of the parameters which best estimates the experimental data. We realize that a single parameter combination may not be adequate in describing the experimental data; instead, there exists a parameter hyper-space adequately satisfying the data. Hence an ensemble parameter estimation was performed by randomly generating initial guesses from the hyper-space of the sensitive parameters. The model was simulated with 10000 random parameter samples; for each of these simulations the least square estimate is determined between

experimental data and model output. These simulations were run for each mechanism and condition under investigation, and parameter ensembles were generated by considering only those parameter sets which meet certain error constraints. Following the above detailed methodology we evaluated the model predictions obtained from the different mechanisms and compared them with experimental data to determine the most plausible mechanism.

5.3 RESULTS

5.3.1 Experimental data

The hESC culture was analyzed for cellular growth and death dynamics during endoderm induction by both Activin A (Condition A) and Activin A/FGF2/BMP4 (Condition B) conditions as illustrated in Figure 5.2. Cellular growth kinetics was found to exhibit nonlinear dynamics, while cell death remained predominantly linear over time. A proliferation lag time is exhibited in Condition A up until Day 3, during which time the number of live cells decreases because of cell death. Beyond this time, cells begin to proliferate in a roughly linear fashion. Interestingly, the majority of cell growth in Condition B occurs before Day 3.

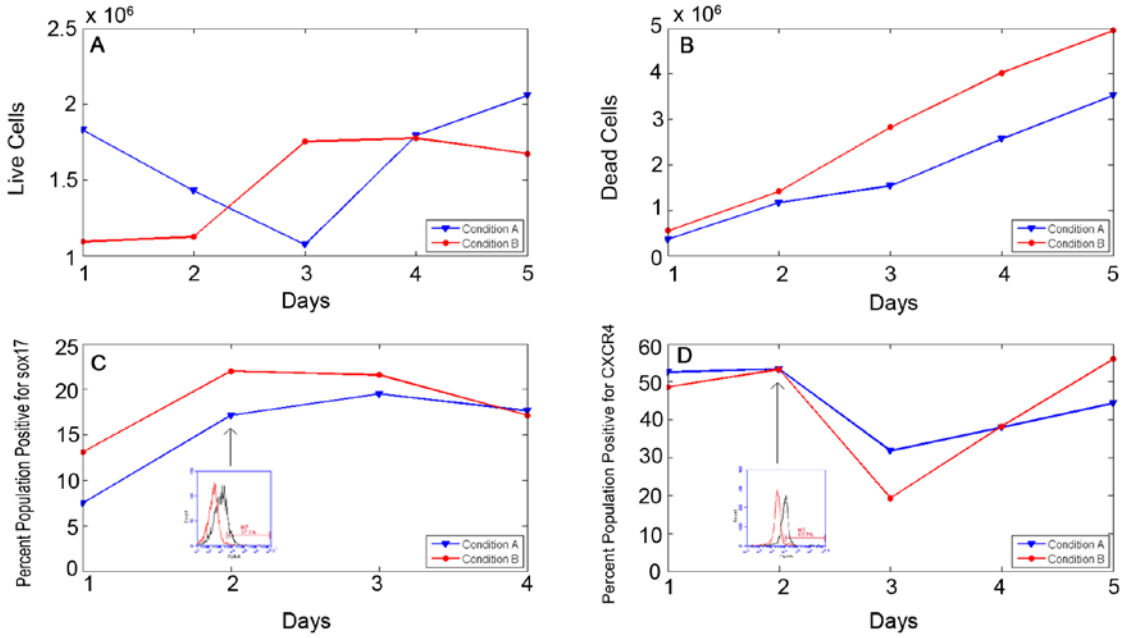

Figure 5.2. Experimental results of cell behavior during endoderm induction

Cellular growth (a) and death (b) dynamics for Conditions A and B. Temporal behavior of cellular population positive for Sox17 (c) and CXCR4 (d). Inset: Representative output of flow cytometry data. Red histogram: secondary antibody only control. Black histogram: sample. Red gate denotes sample taken to be positive.

The differentiated cell population was analyzed by flow cytometry for percent of cells positive for Sox17 and CXCR4 for each day of differentiation. Figure 5.2 represents the dynamics of Sox17 (c) and CXCR4 (D) expression for both the experimental conditions, with the insets illustrating representative flow cytometry results. Details of the flow data are presented in Figure A3. The population positive for Sox17 exhibits a maximum percentage of ~19-23% at ~2-3 days (depending on the condition), with subsequent decay. The fraction of cells positive for CXCR4 is relatively constant until the second day, after which there is a significant drop. Subsequently, there is an approximately linear increase in the CXCR4 population which is more prominent in Condition B.

5.3.2 Mathematical model

5.3.2.1 Model parameter analysis

In the next step the developed stochastic model was used to test the proposed mechanisms for agreement with experimental data. The mathematical model involves multiple parameters which require detailed analysis before the model can be used for prediction. The parameters can be grouped into two categories: (a) simulation parameters which affect the convergence behavior of the simulation and (b) model parameters which affect specific model output for the converged simulation. Two parameters were identified to be simulation parameters: initial cell population and number of stochastic runs. These parameter values were optimized by performing a two-dimensional convergence study, as illustrated in Figure 5.3 for the CXCR4-positive population output. Overall it is observed that the initial cell population has a more dominant effect on convergence, while the effect of stochastic runs was rather weak beyond 2000 runs. Following this analysis an initial cell population of 9000 and total stochastic runs of 4000 was used for subsequent simulations.

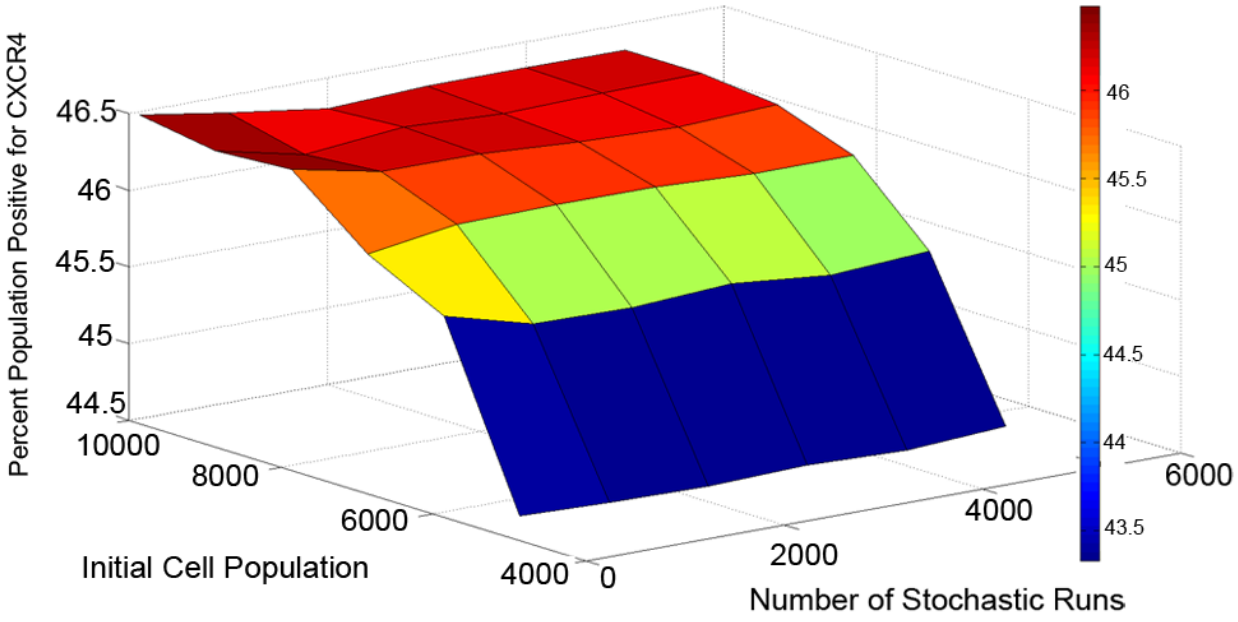

Figure 5.3. Convergence study of simulated cell population over various initial cell populations and total stochastic runs

Output is percent of the simulated population positive for CXCR4, averaged over all stochastic runs at Day 5.

A detailed sensitivity analysis was performed for the model parameters in order to determine the relative importance of the parameters in determining model output. As detailed in Equation 5.1, the measure of histogram distance is used to represent the parameter sensitivity associated with a specific model output. Figure 5.4(a) illustrates the model parameter sensitivity to the output of cellular growth, as concluded from the shift in histogram distance (inset). A clear jump in the sensitivity is observed, with a large difference between parameters with low and high sensitivity.

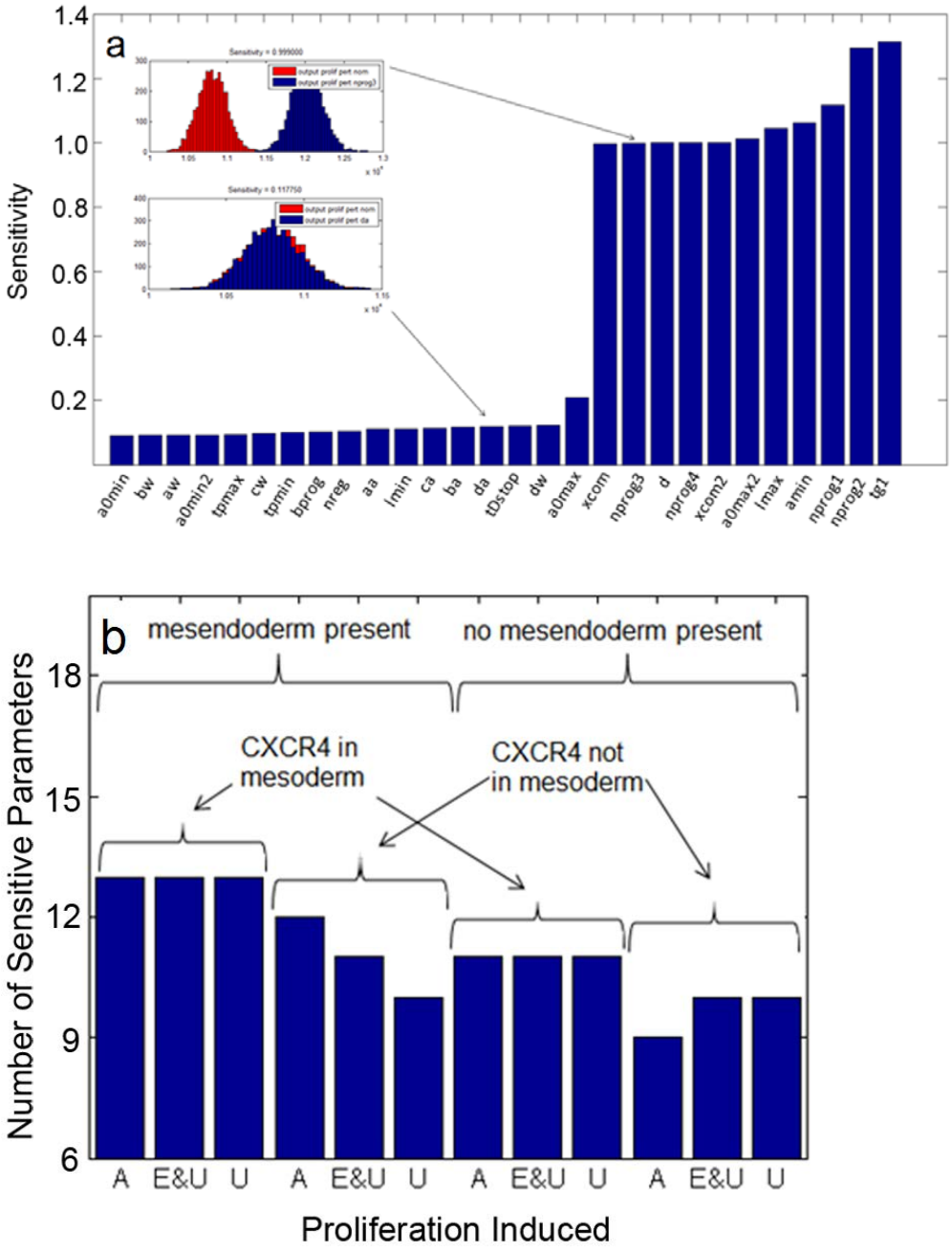

Figure 5.4. Sensitivity analysis of population based model

(a) Cellular growth sensitivity to each of the parameters, perturbed by 10%. Parameter definitions listed in Table A6. (inset) Comparison of cellular growth output histogram from nominal (red) and perturbed (blue) parameters. (b) Number of sensitive parameters determined for each mechanistic model. Proliferation induced: A, all phenotypes; E&U, endoderm and uncommitted (hESC and mesendoderm); U, uncommitted only.

While Figure 5.4(a) represents the parameter sensitivity to model output of cellular growth, similar analysis was also performed for all of the model outputs. Overall it was observed that even though the magnitude of sensitivity differs between outputs, the highly sensitive parameters were mostly conserved across outputs. Furthermore, in the present study we are investigating multiple competing mechanisms, which require modification of the model formulation. Since the effect of such modifications on the parameter sensitivity is not intuitively obvious, similar analysis was repeated for each of the 12 proposed mechanisms. Figure 5.4(b) summarizes the number of sensitive parameters for each of the mechanisms. From the analysis, eight classes of parameters were consistently observed to have the highest sensitivity:

- a_{min}: 'a' value threshold beyond which a cell enters the proliferation phase
- a_{0max}: The initial cell population is randomly assigned an 'a' value, with an upper limit of a_{0max}
- x_{com}: lineage propensity threshold beyond which a cell is committed to a particular lineage
- d: factor with which 'a' decreases
- t_{g1}: time cell stays in G1 phase of cell cycle. Only in this phase can a cell differentiate and transfer to the A regime from the Ω regime
- l_{max}: upper value of range of cell population's life span
- $nprog_i$: factor in determining magnitude of propensity updates for each lineage i
- aa: parameter in determining the probability of a cell transferring from the Ω regime to the A regime

5.3.2.2 Ensemble parameter estimation

Having determined the sensitive parameters for each of the mechanisms, the next step is to determine the optimum parameter values which will result in best agreement with experimental data. In literature, biological samples are typically defined as being 'sloppy' [76] with a broad ensemble of parameters satisfying the error constraints. Accordingly we also target identification of representative ensemble of parameters. The model is formulated to capture the dynamics of cellular growth, death and differentiation, the output of differentiation being of most interest. Hence the model parameters were optimized with respect to differentiation dynamics, while growth kinetics and the dynamics of cell death were used for verification. A projection of the simulated error onto a 2-dimensional parameter space (for the mechanism which incorporates mesendoderm and promotes proliferation of both uncommitted and endodermal cells without CXCR4 being expressed in mesoderm ('Mechanism B')) is shown in Figure 5.5(a). Although it was initially thought that a trend might be observed between the error and values of the parameter ensemble, Figure 5.5(a) shows that there is a lack of any correlation between multiple parameters and associated errors (shown for parameter 'd'; further analysis of all parameter combinations yielded the same result). Figure 5.5(b) illustrates the minimum ensemble error for each of the proposed mechanisms simulated under the two endoderm induction conditions, the error being evaluated according to least square formulation.

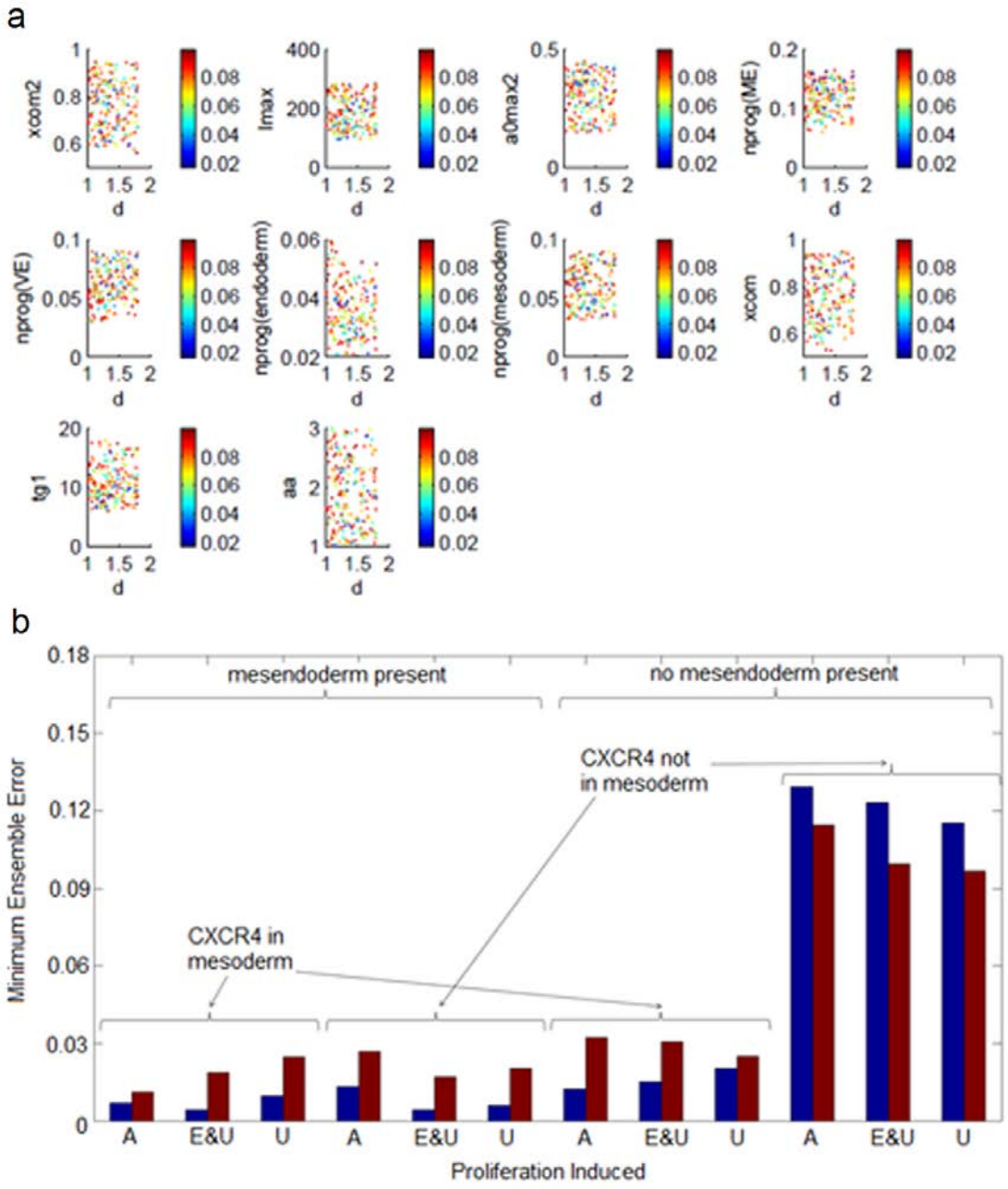

Figure 5.5. Ensemble parameter estimation and model errors

(a) Parameter values for Mechanism B ensemble yielding errors of less than 0.025. Each parameter is compared to the most sensitive parameter, 'd'. Color bar denotes the ensemble error for that particular parameter value. (b) Minimum ensemble error generated for each mechanistic model. Proliferation induced: A, all phenotypes; E&U, endoderm and uncommitted (hESC and mesendoderm); U, uncommitted only. Blue, Condition A; Red, Condition B.

5.3.2.3 Mechanism evaluation: endoderm induction by Activin A

As shown in Figure 5.5(b), the absence of mesendoderm gives rise to large errors, in some cases an order of magnitude higher than their counterpart models which include mesendoderm. If one considers the Activin A only condition, the most accurate mechanisms include those which incorporate mesendoderm and promote proliferation of both uncommitted and endoderm germ layer both with (‘Mechanism A’) and without (‘Mechanism B’) CXCR4 in mesoderm. Since Figure 5.5(b) illustrates the accuracy of the model in predicting differentiation dynamics only, the performance of the 2 prospective mechanisms were further verified with the help of growth kinetics and the dynamics of cell death. Figure 5.6 illustrates the ensemble simulation of all the model outputs and its comparison with experimental data. While both the mechanisms had excellent performance in predicting Sox17 and CXCR4 dynamics (Figure 5.6 (c,d,g,h) they differed significantly in predicting growth kinetics and cell death dynamics (Figure 5.7 (a,b,e,f). Figure 5.6 clearly illustrates that Mechanism B performs better in describing both growth kinetics and cell death dynamics compared to Mechanism A. Hence the former was chosen to be the most likely mechanism for Condition A.

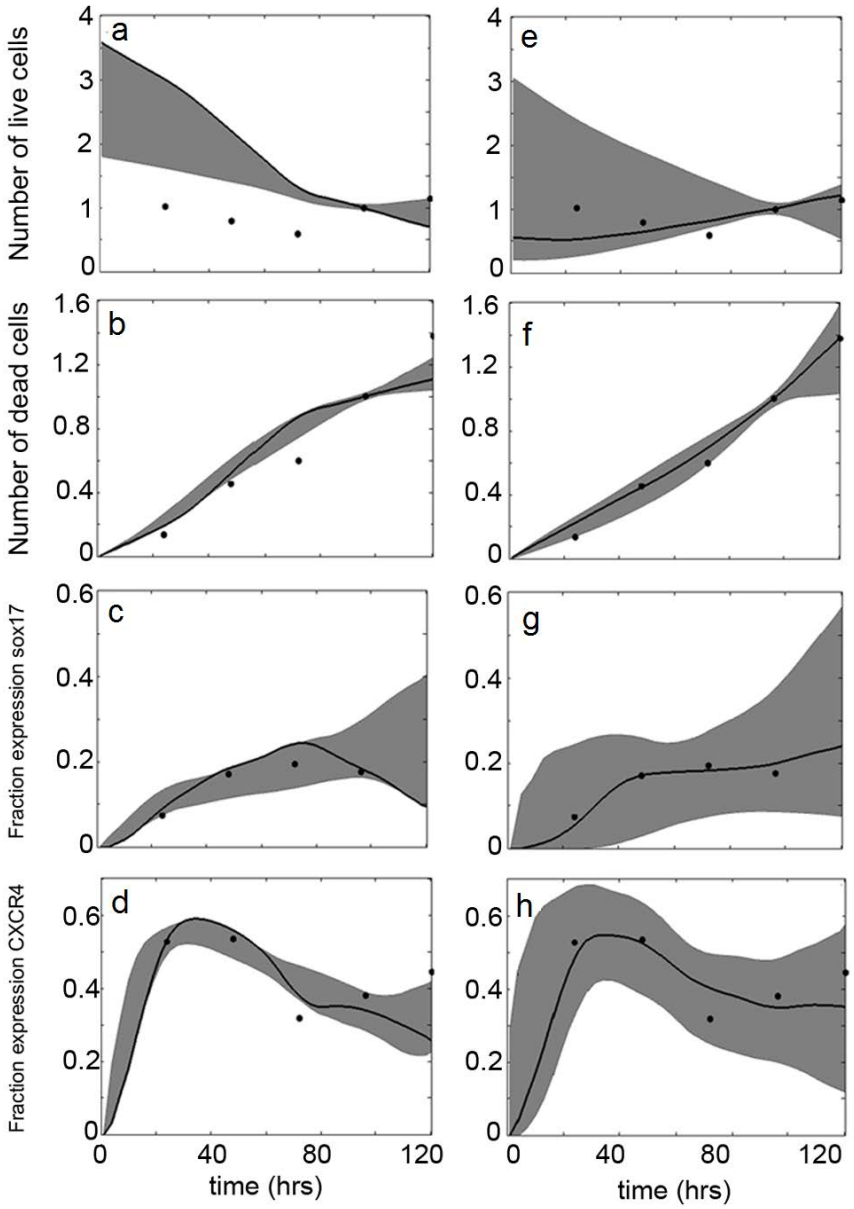

Figure 5.6. Simulated output dynamics compared to experimental data (Condition A)

Grey band denotes the ensemble of simulations having an error less than the threshold, with the single solid black curve showing the best fit. Black circles represent the experimental data points. (a-d): Growth kinetics, cell death, fraction of population positive for Sox17 and CXCR4 dynamics, respectively, of Mechanism A; error threshold of 0.05. (e-f) Growth kinetics, cell death, fraction of population positive for Sox17 and CXCR4 dynamics, respectively, of Mechanism B; error threshold of 0.025. Cellular growth and death normalized to Day 4.

5.3.2.4 Mechanism evaluation: endoderm induction by Activin A supplemented by growth factors

Condition B (Activin A supplemented with FGF2 and BMP4) proved more difficult to describe via the investigated mechanisms, mainly because CXCR4 dynamics exhibits a faster and more prominent drop as compared to Activin A only condition. As shown in Figure 5.5(b), the two mechanisms which give lowest error for Condition B are the ones which incorporate mesendoderm, have CXCR4 present in mesoderm and promote proliferation of all phenotypes (‘Mechanism C’) and the previously described Mechanism B. The simulated dynamics of these two mechanisms with experimental data of Condition B are shown in Figure 5.7. As with Condition A, although the incorporation of CXCR4 in mesoderm gives a small error, the simulated profiles of growth kinetics and cell death are not in agreement with experimental data. The next best mechanism, Mechanism B, exhibits both a low error and good results with all outputs, so it was again chosen as the most likely mechanism.

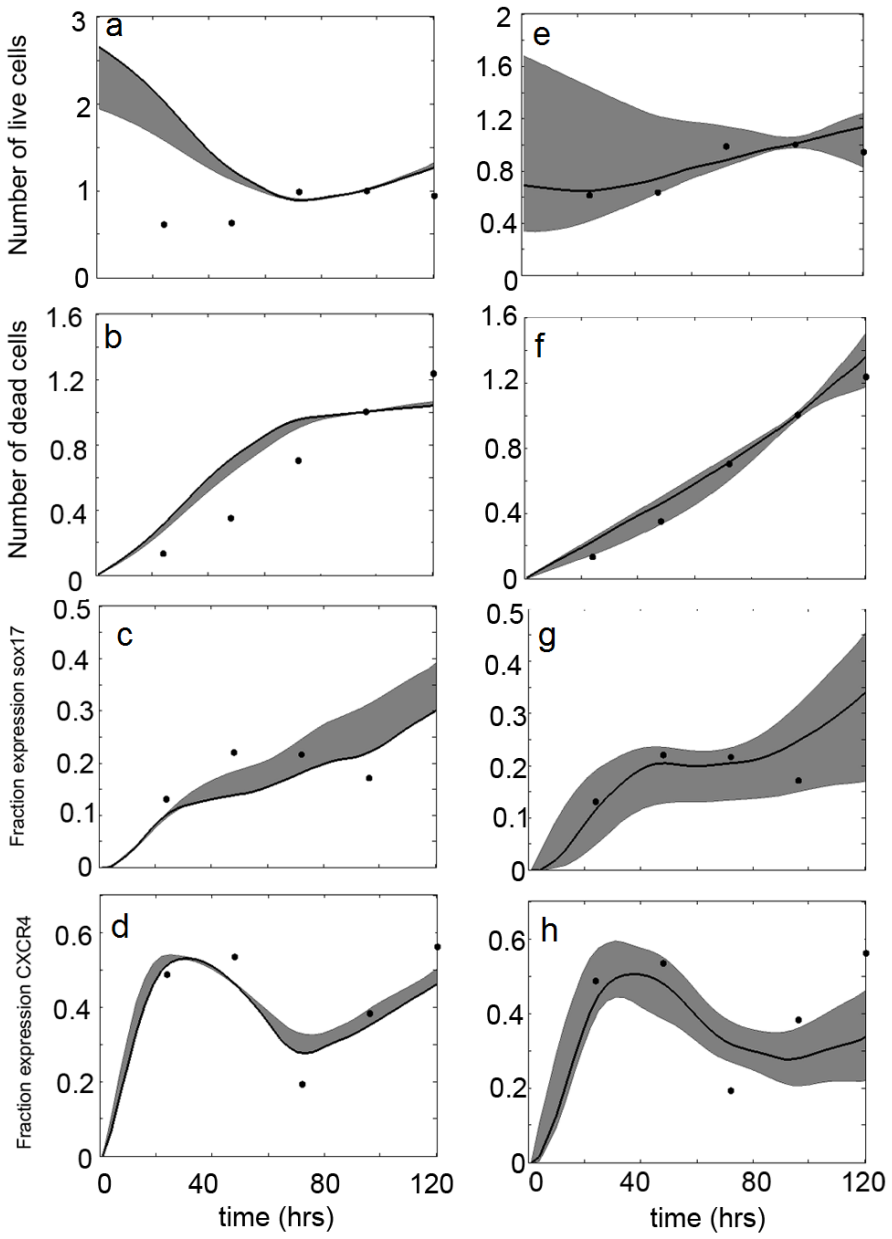

Figure 5.7. Simulated output dynamics compared to experimental data (Condition B)

Grey band denotes the ensemble of simulations having an error less than the threshold, with the single solid black curve showing the best fit. Black circles represent the experimental data points. (A-D): Growth kinetics, cell death, fraction of population positive for Sox17 and CXCR4 dynamics, respectively, of Mechanism C; error threshold of 0.1. (E-F) Growth kinetics, cell death, fraction of population positive for Sox17 and CXCR4 dynamics, respectively, of Mechanism B; error threshold of 0.025. Cellular growth and death normalized to Day 4.

It is therefore reasonable to conclude that during endoderm induction with the conditions described above, undifferentiated stem cells first differentiate into a mesendoderm germ layer with subsequent differentiation to endoderm and mesoderm, the latter not expressing CXCR4. Furthermore, the induction condition seems to promote proliferation of both pluripotent and endoderm-like cells. The optimized parameters of this mechanism are shown in Table 5.1, with definitions of parameters in Table A6.

Table 5.1. Comparison of the best fit parameter set between the two conditions

Parameter	Condition A	Condition B
a0max2	0.3	0.161
xcom	0.8	0.767
xcom2	0.9	0.79
d	1.2	1.32
tg1	12	14.6
lmax	190	222
nprog(ME)	0.11	0.0879
nprog(endoderm)	0.04	0.0278
nprog(mesoderm)	0.06	0.0457
nprog(VE)	0.06	0.0611
aa	2	2.07

5.3.3 Model validation

The power of mathematical models lies in their predictive capacity. The predictive capacity of our proposed model was thus tested by simulating the population dynamics of cell types for which no *a priori* data was used in constructing the model. The chosen populations were that of undifferentiated cells and mesendoderm cells. The simulated profile of the

undifferentiated cells (Figure 5.8(a) and (b)) shows an exponential decay to a final value of 10% of the cellular population. This final value was reached in approximately 3 days. The mesendoderm cell population was predicted to display more interesting dynamics, with a transient increase in cell population over the first day, followed by a decreasing trend over the next few days (Figure 5.8(d) and (e)).

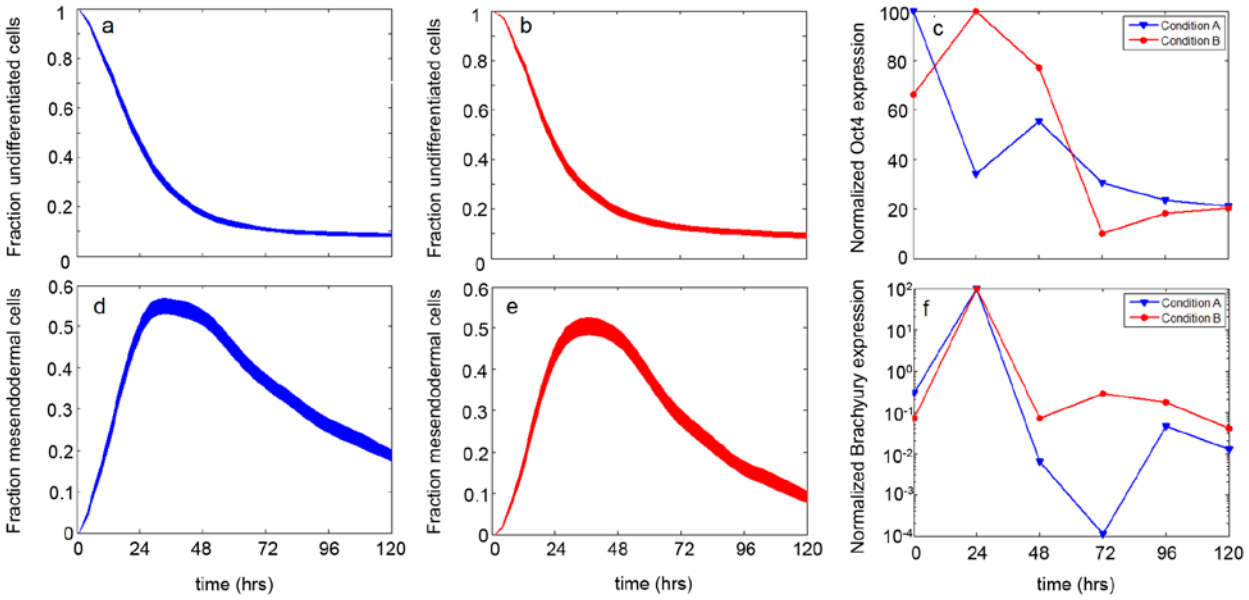

Figure 5.8. Validation of model with experimental gene expression data

Simulated dynamics of the undifferentiated (A (Condition A), B (Condition B)) and mesendoderm (D (Condition A), E (Condition B)) phenotypes were compared to experimental data of their respective genes, measured by qPCR: Oct4 (Undifferentiated; C) and Brachyury (Mesendoderm; F). The simulated dynamics bands represent 4000 stochastic simulations using the optimized parameters of Mechanism B. mRNA levels were measured with time using qPCR. Data was first normalized to the housekeeping gene Gapdh then to undifferentiated cells. Fold change levels, determined by the $2^{-\Delta\Delta Ct}$ method, were then normalized to the maximum level for each respective gene (data reported as percent of maximum fold change).

In the next step the validity of such prediction was verified by conducting further experiments to analyze the dynamics of undifferentiated cells by Oct4 gene expression and that of mesendoderm cells by Brachyury expression. While the comparison of population dynamics with mRNA levels is not exact, under the assumption of efficient translation they become comparable. Figure 5.8 illustrates the comparison of experimental data with model prediction, which are found to have excellent agreement given that the model was generated with no information of these specific cellular dynamics. Oct4 levels exhibit a decay to a final value of around 20% of the maximum, at a time which correlates with simulated predictions (3 days). Brachyury levels showed a similar trend as was predicted by the model. It reached a maximum around 24 hrs, following which it gradually decayed over time.

5.4 DISCUSSION

5.4.1 Mechanism alternatives and identification

Definitive endoderm was induced in hESC through two different pathways: the addition of Activin A and Activin A supplemented with FGF2 and BMP4. The population-based model, revised from the model originally developed for hematopoietic stem cells [37, 183], tracks individual cell behavior based on a number of set rules. The focal point of the rules is lineage propensity updating, wherein the likelihood of differentiation to a particular lineage is stochastically updated per time step. The lineages to which hESC can differentiate are definitive endoderm, visceral endoderm, and mesoderm. Depending on the specific mechanism of the model, hESC can first give rise to visceral endoderm and mesendoderm, with the latter

differentiating into definitive endoderm and mesoderm. In the current model, the ectoderm germ layer has been omitted. From previous literature [188], hESC induced towards endoderm show low expression of ectoderm markers (Sox1). Commitment levels to ectoderm would be low, and therefore adding the additional ectoderm lineage would not enhance the model.

It is important to note that the nonlinearity observed in the differentiation dynamics contributed significantly towards identification of a robust mechanism. Sloppiness of biological parameters is well reported [76] with ranges of values being large and sensitivities between parameters varying considerably; this can make robust mechanism identification challenging. Quite interestingly, the observed dynamics of the presented study could only be explained by a single specific mechanism. Even a rigorous search of the parameter hyperspace did not yield an alternate potential mechanism. Regarding the nonlinearity of CXCR4 expression, two possible explanations can be: (1) mesendoderm, expressing CXCR4, further differentiates to phenotypes which might not express the surface protein; and (2) the cellular environment might promote a higher rate of death of a certain cell phenotype which expresses CXCR4. These dynamics, along with those of Sox17, proliferation, and cell death, led us to investigate a total of 12 possible mechanisms. The majority of the mechanisms investigated was unable to capture the temporal behavior of these outputs, and therefore was discarded. The only mechanism which is able to accurately explain the experimental dynamics is one which does not have mesoderm expressing CXCR4, incorporates mesendoderm, and promotes proliferation of hESC and the mesendoderm and endoderm germ layers. This proposed mechanism is shown in Figure 5.9.

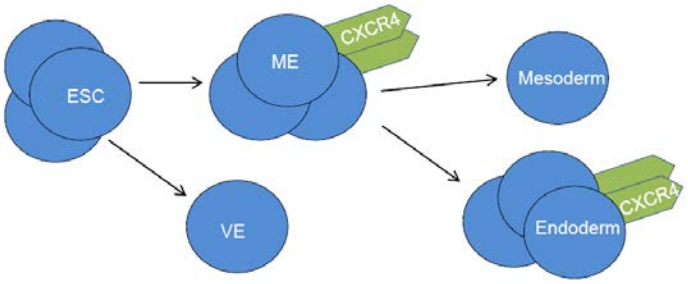

Figure 5.9. Proposed differentiation scheme of hESC during endoderm induction as generated by the population-based model

Shown is the presence of mesendoderm, the lack of CXCR4 in mesoderm, and selective phenotype proliferation. ME: mesendoderm, VE: visceral endoderm

One of the purposes of the present study was to investigate several aspects of differentiation which have faced conflicting reports in the past and to offer further insight using a mathematical analysis. One of these features is the presence of surface receptor, CXCR4. McGrath *et al*.[182] and Yusuf *et al*. [189] have reported that embryonic mesoderm expresses CXCR4 *in vivo*, depending on the stage of embryo development, whereas Takenaga *et al.* [190] reports using CXCR4 as a definitive endoderm marker with other markers used for mesoderm. Our results indicate that although both possibilities give low error with respect to Sox17 and CXCR4 population dynamics (depending on which phenotype proliferation is induced), only when CXCR4 is absent in mesoderm do we obtain qualitative agreement in the cellular growth and death temporal behavior. Furthermore, the majority of studies which follow embryo development *in vivo* or differentiation of ESC *in vitro* (for example [191-193]) include the mesendoderm as an intermediate phenotype arising from the differentiation of ESC which subsequently differentiates to endoderm and mesoderm rather than considering the latter two phenotypes differentiating directly from ESC. The model developed in the current study indeed

comes to the same conclusion: the mesendoderm germ layer needs to be considered in order to accurately describe experimental dynamics.

5.4.2 Comparison between two differentiation conditions and parameter estimation

The endoderm induction of hESC was conducted under two different conditions with the objective of investigating mechanistic differences between these two pathways. Quite interestingly, both conditions could be explained by the same, single mechanism, while the rejected mechanisms failed to describe the dynamics even after a thorough search of the parameter space. However, there were significant differences in optimum parameter values. One prominent difference between the two conditions was their differentiation potential after being committed to the mesendoderm germ layer. 'a0max2' is lower for Condition B, indicating that mesendodermal cells will more quickly reach the pro-differentiation and –proliferation regimes. This is also evident from the higher level of 'd', although this is for both stages of differentiation. Also, cell commitment for Condition B can be considered expedited when considering the lower value of 'xcom2', which is the propensity threshold beyond which a mesendodermal cell is considered committed to either endoderm or mesoderm. Therefore, Activin A supplemented with FGF2 and BMP4 drives differentiation towards endoderm/mesoderm to a higher degree than Activin A alone.

As detailed earlier, the developed model for the optimum mechanism could accurately capture the experimentally observed dynamics of differentiation. Quite interestingly, only a single mechanism could adequately describe the experimental data, while the rejected mechanisms failed to describe the dynamics even after a thorough search of the parameter space.

5.5 CONCLUSIONS

The objective of the work in this chapter was to investigate the mechanism of differentiation of hESC during endoderm induction while capturing cell-to-cell variability through an integrated experimental and mathematical approach. We experimentally determined the dynamics of differentiation upon endoderm induction of hESC and use these data along with a population-based stochastic model to determine the mechanisms of differentiation. The model can track growth kinetics, the dynamics of cell death, and the dynamics of differentiation into the germ layers. Comparison of these simulated outputs with experimental data enables determination of the dominant mechanism of differentiation. Not every model alternative captured experimental dynamics; indeed, only one mechanism was able to describe the protein, cell growth and cell death dynamics. This mechanism indicates that during endoderm induction, certain phenotypes display increased proliferation-endoderm and uncommitted cells. Furthermore, several behaviors during initial differentiation that have received conflicting experimental results in literature have been predicted: the presence of an intermediate mesendoderm germ layer which arises before commitment to endoderm and mesoderm; and CXCR4, a surface protein, is present on the mesendoderm and endoderm germ layers, but not on mesoderm. The model and predicted mechanism was validated against additional experimental observations of the temporal behavior of specific cellular phenotypes. Even though these data were not used to build the model, the model performed extremely well in capturing their dynamics.

6.0 OVERALL CONCLUSIONS AND FUTURE WORK

In this work, we have analyzed the self-renewal and differentiation of the embryonic stem cell system at three different levels, and demonstrated the utility of computational modeling and mathematical analyses in extracting information that would otherwise be difficult from a purely experimental approach. Specifically, we have focused on the intracellular gene networks, extracellular substrate properties, and population dynamics of the ESC system, concentrating on the heterogeneity and complexity of each system.

We have integrated a mixed-integer optimization algorithm with the bootstrapping technique to predict intracellular robust gene regulatory network interactions in the presence of high levels of transcriptional noise. Our developed algorithm is able to accurately predict gene interactions and associated strengths from gene expression dynamics. In addition to intracellular analysis, we developed a statistical approach to determine the effects of complex extracellular cues on ESC behavior. Using this approach we identified specific microstructural features contributing to spontaneous differentiation and gene expression patterning. Finally, to capture population dynamics and heterogeneity, we have built a stochastic model which describes cell cycle transition. We extend this model to capture differentiation during germ layer commitment. We utilize both models to make informed predictions on the mechanisms of differentiation and self-renewal. At these three levels, we have demonstrated the utility of mathematical models in information extraction, especially when variability is an issue. Future work will focus on

utilizing these methods and results to improve directed differentiation towards insulin producing β cells for diabetes treatment. Specifically, Aims 1 and 2 will be used for practical improvement of the differentiation protocol, upregulating genes dictating mature pancreatic development, including *PDX1, NGN3,* and *INSULIN*. Aim 3 will focus on characterizing cells during the maturation process, including tumorigenicity potential and differentiation status.

6.1 IDENTIFICATION OF ROBUST INTRACELLULAR GENE REGULATORY NETWORKS

Our first aim focused on the advancement of algorithms to reverse engineer gene regulatory networks. We have developed a network identification algorithm to accurately infer both the topology and strength of regulatory interactions from time series gene expression data in the presence of significant experimental noise and nonlinear behavior. In this novel formulism, we have addressed data variability in biological systems by integrating network identification with the bootstrap resampling technique, hence predicting robust interactions from limited experimental replicates subjected to noise. Furthermore, we have incorporated nonlinearity in gene dynamics using the S-system formulation. The basic network identification formulation exploits the trait of sparsity of gene regulatory interactions. Towards that, the identification algorithm is formulated as an integer-programming problem by introducing binary variables for each network component. The objective function is targeted to minimize the network connections subjected to the constraint of maximal agreement between the experimental and predicted gene dynamics. The developed algorithm is validated using both *in silico* and

experimental data sets. These studies show that the algorithm can accurately predict the topology and connection strength of the *in silico* networks, as quantified by high precision and recall, and small discrepancy between the actual and predicted kinetic parameters. Furthermore, in both the *in silico* and experimental case studies, the predicted gene expression profiles are in very close agreement with the dynamics of the input data.

This algorithm is general in nature, meaning that it can be applied to any biological system for which time series gene expression data is available. Therefore, while the formulation has been validated against *in silico* and *E. Coli* data, its formulation, biological relevancy, and results are applicable to any gene regulatory network. This algorithm can therefore be directly utilized for the identification of regulatory networks governing ESC differentiation. As the primary goal of our research is to differentiate hESC to mature, insulin producing β-cells for diabetes treatment, determining the gene connections involved in pancreatic development is paramount. While population models, signaling models, and the like have the ability to offer mechanistic information and predictive capability, without understanding the effects in the nucleus and how genes are transcribed there is always an amount of speculation in the effects of system perturbations. With detailed information on the gene regulatory network, one can obtain specific information on how genes in the differentiation pathway behave. While information on how transcription factors interact is somewhat available, this information comes primarily from developmental *in vivo* studies, and detailed interaction information is rarely available for hESC in culture. Furthermore, much of the interaction information is qualitative, and quantitative kinetic behavior is lacking. The current algorithm can aid in this understanding.

Researchers have shown that gene regulatory networks involved in the developmental process act as modules, in that for each developmental stage there is a central gene as a master

regulatory, and that the connections between it and the other genes in the network are minimal (hence the formulation of sparsity in the current algorithm is directly applicable). Connections between hubs of different developmental stages are also minimal [194]. In future work these guiding features can help identify the regulatory modules and associated connections during *in vitro* hESC differentiation towards β-cell development. Preliminary work should first focus on compiling, from literature, the genes involved in *in vivo* endoderm, pancreatic progenitor, and endocrine progenitor/β-cell maturation development. Experimentally, hESC will be differentiated to these phenotypes, and the dynamics of the genes identified in the literature search will be quantified at each stage using qPCR. Having this temporal behavior, the network identification algorithm described herein can be implemented, and network topology and strength identified. While hESC gene expression is often characterized by high levels of variability, the bootstrapping approach utilized in the algorithm will be able to identify the robust connections in the presence of the replicate noise.

By obtaining the information on the gene regulatory networks for these differentiation stages, and how the nodes interact with each other, a much more thorough understanding of hESC differentiation can be obtained, and can guide directed differentiation towards β-cells. For instance, from our previous differentiation studies, we have observed that at the pancreatic progenitor stage the marker indicative for this stage, PDX1, is expressed to a high level (~80% of cells positive for the protein). However, the final maturation stage yields relatively low amounts of insulin (10-20%). Using the information obtained by the gene regulatory network algorithm, we hope to improve this. For example, we hope to determine how PDX1 interacts with the early developmental genes in the endocrine node (NGN3) and subsequent β-cell node (INSULIN). Knowing these connections, we may be able to gauge which gene(s) should be perturbed to gain

larger upregulation of NGN3 and INSULIN. Once these genes are known, their expression levels will be experimentally altered. The method of this alternation will be determined from the algorithm results, e.g. if a gene should be up-regulated, down-regulated, or silenced. These gene perturbations can be experimentally done through a wide-range of techniques, including siRNA and gene knockouts. Growth factors can also be used to change the expression levels of the target genes if the pathways relating the growth factor to the target gene are known.

6.2 IDENTIFICATION OF EXTRACELLULAR SUBSTRATE CUES INFLUENCING DIFFERENTIATION

Our second aim focused on quantifying extracellular factors and determining which specific substrate features effect ESC differentiation behavior. The system we investigated was that of mESC on various fibrin gels. We synthesized a range of fibrin gels by varying fibrinogen and thrombin concentrations, which led to a range of substrate stiffness and microstructure. mESC were cultured on each of these gels, and characterization of the differentiated cells revealed a strong influence of substrate modulation on gene expression patterning. To identify specific substrate features influencing differentiation, the substrate microstructure was quantified by image analysis and correlated with stem cell gene expression patterns using a statistical model. Overall, it was shown that for the presented class of fibrous substrates, the contribution of substrate microstructure to mESC differentiation was stronger than that of macroscopic stiffness. Furthermore, it was found that fibrous microstructural features had a strong influence on ESC differentiation to endoderm lineage, with fiber alignment being the most influential.

The approach applied to and results obtained from the system of mESC on fibrin gels may aid in the design of materials with a preferred microstructure to more effectively guide pancreatic specific stem cell differentiation with efficient transplantation capabilities. For eventual transplant of ESC-derived cells, appropriate scaffolds will have to be used. Because of fibrin's characteristics (described in Chapter 3), this natural material could be an excellent choice for a transplantation scaffold. As a next step, this scaffold should support the phenotype of interest. In our case, this is β-cells. Therefore, if fibrin could be used to direct differentiation to this phenotype, transplantation potential could be greatly enhanced. The results in Chapter 3 show that endoderm differentiation, a precursor to β-cells, is sensitive to the microstructural features of fibrin, and that certain features are more influential than others. The next step would be to utilize this information to optimize the values of these features in order to optimize endoderm gene upregulation. Specifically, fiber alignment was shown to have the greatest influence on endoderm differentiation. Therefore, fibrin gel fiber alignment will be modified to improve endoderm differentiation. To determine optimal fiber alignment values, a more precise response surface of the differentiation patterning to the alignment will be generated through design of experiments. Furthermore, to isolate the effects of the alignment, the other fibrous properties should be kept invariant. However, this would be difficult to achieve using the current protocol for fibrin fabrication. An attractive alternative could be to create these fibrin gels with electrospinning, which allows for much more controllable features. The effectiveness of this method has been demonstrated with fibrin [114]; this method may therefore have the potential to create fibrin gels with precise fiber alignment values. Once this is accomplished, the best fit equations of fiber alignment vs. endoderm gene expression will be optimized to determine what values will yield the largest endoderm upregulation. Again using electrospinning, fibrin gels will

be created with this optimized fiber alignment, with endoderm being upregulated. In addition to endoderm, the methodology can be applied to further stages of differentiation. In this way, fibrin substrates can be designed with optimal microstructural topology to guide differentiation through pancreatic lineage, with the resulting cell being ready for transplantation on the fibrin scaffolds.

6.3 HETEROGENEOUS POPULATION DYNAMICS OF HESC CELL CYCLE AND DIFFERENTIATION

The final objective of this work was to capture emergent population behavior from single cell behavior while utilizing population dynamics to gain mechanistic insight. This model was first applied to the embryonic stem cell cycle. The developed cell cycle model was compared to experimental data, using agreement with these dynamics to postulate on prominent mechanisms. These results allow us to theorize that G1 times are tightly controlled during pluripotency, but gradually become more variable with increased maturity, and that the population distributions of the phase residence times are governed by the differentiation induction condition acting in different ways at the single cell level.

These proposed behaviors can be validated experimentally in several ways. As shown in Figure 4.5 (b,c), mean cell cycle and phase residence times can be validated by various assays, including cell dilution dyes and DNA stains coupled with synchronization. These validation experiments were in excellent agreement with the model predictions. Further population and ensemble model predictions can be validated with more sophisticated techniques. The most powerful of these is single cell tracking. By following individual cells with time-lapse

microscopy, and determining the intermitotic times and G1 times of individual cells (the latter of which being determined through the FUCCI reporter), the cell cycle and G1 residence time distributions and variabilities can be determined and used to validate model predictions. Single cell tracking can also be used to validate the G1 lengthening mechanisms. One model prediction suggests that G1 lengthening is exponential in nature. Again, single cell tracking coupled with the FUCCI reporter can quantify the exact G1 time, and capture how this changes with time and cell cycles, thereby offering a technique for model validation.

An area of future work which may lend itself to the field and which was not considered in the current model is how cells interact with one another. This could be accomplished by incorporating the current model into an agent based model (ABM) platform. ABM could capture spatial orientation of the cells and the effects of cell-cell interactions. While cell-cell interactions are undoubtedly important in governing cell behavior, they may also be important specifically when considering the cell cycle. Sela *et al.* has reported that high cell density reduces differentiation in the G1 phase, and that this reduction is enhanced when the surrounding cells are in the S and G2/M phases [142]. This effect could be captured by ABM, and further aid in mechanistic understand and hypothesis testing.

The cell cycle model will be used for precise characterization of the maturing ESC and ESC-derived progenitor cells. Specifically, the model can characterize which cells and cellular populations have fast or slow proliferation rates, which is relatable to tumorgigenicity, which is a challenge in using ESC for therapeutic use [195]. Quantifying tumorigenicity during the differentiation process is a challenging experimental task. Our model allows for the dynamics of cell cycle intermitotic times to be tracked, thereby determining possible origins of tumorigenicity. Our results suggest that with differentiation towards pancreatic progenitors there

exists a subpopulation of cells with faster division times. This analysis will be extended to the end maturation to determine if this subpopulation is still prominent. If so, then the differentiation protocol will need to be modified in order to reduce this population before eventual transplantation.

We next extended this population model to focus on the specifics of differentiation, and initial lineage commitment during directed differentiation towards endoderm. The differentiating cell population was analyzed daily for cellular growth, cell death, and expression of the endoderm proteins Sox17 and CXCR4. The stochastic model starts with a population of undifferentiated cells, wherefrom it evolves in time by assigning each cell a propensity to proliferate, die and differentiate using certain user defined rules. Twelve alternate mechanisms which might have describe the observed dynamics were simulated, and an ensemble parameter estimation was performed on each mechanism. A comparison of the quality of agreement of experimental data with simulations for several competing mechanisms led to the identification of one which adequately describes the observed dynamics under both induction conditions. The results indicate that hESC commitment to endoderm occurs through an intermediate mesendoderm germ layer which further differentiates into mesoderm and endoderm, and that during induction proliferation of the endoderm germ layer is promoted. Furthermore, our model suggests that CXCR4 is expressed in mesendoderm and endoderm, but is not expressed in mesoderm. Comparison between the two induction conditions indicates that supplementing FGF2 and BMP4 to Activin A enhances the kinetics of differentiation than Activin A alone.

This model will next be applied to later stages of ESC differentiation. This future work will not only provide mechanistic insight into lineage commitment during late-stage pancreatic specification, but will be used to determine which perturbations to the differentiation protocol

should be focused on to improve yield of more mature cell types. In particular, the population model can be extended to the process of generating endocrine progenitors from pancreatic progenitors, and subsequent derivation of single horomone-expressing cells. A challenge in the field of β-cell differentiation is the yield of insulin positive cells at the final stage of maturity. As mentioned above, in our own work, we see relatively efficient yield of pancreatic progenitors, but low yield (10-20%) of insulin positive cells. Utilizing the population model platform for later stages of differentiation may improve this yield. For instance, what mechanisms are involved when moving from a relatively homogeneous population of pancreatic progenitors to a very heterogeneous population of more mature cells? Is the proliferation rate of the less committed cells overtaking the more mature cells, leading to a dilution effect (as demonstrated in Chapter 4)? Is the efficiency of α-cell differentiation much more than that of β-cells? Furthermore, we have also shown that after maturation, out of the insulin positive cells within the cellular population, there is a significant fraction which is polyhormonal (e.g. expresses glucagon and insulin). These cells are not truly mature, and would be difficult to use clinically. What is the origins of co-expression, and what possible perturbations to the differentiation protocol can reduce this? As with the endodermal stage, we hope to answer these questions by utilizing our model and comparing model alternatives to experimentally derived population dynamics. Once these questions are answered, modifications to the differentiation protocol will be made to improve β cell yield. For instance, if heterogeneity in the cellular population is more prominent at the endocrine progenitor stage, growth factors will be introduced to try and upregulate *NGN3*, rather than focusing on the last stage of maturation.

The process by which embryonic stem cells self-renew and differentiate is complex, and elucidating the mechanisms governing this behavior is difficult using a purely experimental approach. To extract mechanistic information on ESC, we developed diverse modeling techniques which were applied to three levels of the ESC system. Intracellular gene regulatory networks, the primary factor in guiding cellular behavior, are affected by insoluble cues from the extracellular environment, including the cell's associated substrate. Due to the complexity of the substrate and cell-to-cell variability, these cues do not affect all cells equally, leading to a heterogeneous cellular population. Our contribution to the mathematical analysis of each of these cellular levels is helping to improve the mechanistic understanding of the ESC system, which in turn will aid in the efficient scale-up and expansion of pluripotent cells and differentiation towards mature β–like cells. Improvement of these protocols is necessary for ESC to be eventually used for diabetes treatment and other therapeutic applications.

APPENDIX

SUPPLEMENARY TABLES AND FIGURES

Table A.1. Primer sequences used during PCR analysis of differentiation of mESC on fibrin gels

	Gene	Left sequence	Right sequence
house-keeping	β-actin	cagcagttggttggagca	tgggagggtgagggactt
endoderm	sox17	atccaaccagcccactga	acaccacggaggaaatgg
	afp	ctctggcgatgggtgttt	aactggaagggtgggaca
	hnf4	catcgtcaagcctccctct	ccctcagcacacggtttt
	foxa2	gttaaagtatgctgggagccg	cgcccacataggatgacatg
	cxcr4	cgggatgaaaacgtccattt	atgaccaggatcaccaatcca
	gata4	ggcccctcattaagcctcag	caggacctgctggcgtctta
	ttr	ttcacagccaacgactctgg	ggcaagatcctggtcctcct
mesoderm	brach T	aagaacggcaggaggatg	gcgagtctgggtggatgta
	fgf8	acggcaaaggcaaggact	tgaagggcgggtagttga
	gsc	gcaccgcaccatcttca	tcgcttctgtcgtctcca
ectoderm	nestin	ggaggatgtggtggaggat	ttcccgtctgctctggtt
	fgf5	ttcaagcagtccgagcaa	taggcacagcagagggatg
	bmp4	atctggtctccgtccctga	cgctccgaatggcacta
pluripotency	rex1	aaggtcatccacggcaca	tgggagtcatcgcttggt
	oct4	ggagaagtgggtggaggaa	gctgattggcgatgtgag
	sox2	ctggactgcgaactggaga	ttggatgggattggtggt

Table A.2. Regression significance for microcharacteristic relationship

p-values for the significant ($p \leq 0.05$) 2nd order polynomial regressions relating two microstructural features to gene expression, 3D condition

gene	feature 1	feature 2	regression p-value
oct4	node density	fiber length	0.0417
oct4	node density	fiber OI	0.0295
fgf8	pore size	fiber OI	0.0049
fgf8	porosity	connectivity	0.0108
fgf8	connectivity	fiber OI	0.0143
fgf5	fiber length	connectivity	0.0367
fgf5	connectivity	fiber OI	0.0057
nestin	pore size	connectivity	0.0175
nestin	porosity	connectivity	0.014
nestin	fiber length	fiber OI	0.0117
nestin	connectivity	fiber OI	0.0078
sox17	pore size	fiber OI	0.0259
sox17	fiber diameter	fiber length	0.0246
sox17	porosity	fiber OI	0.0246
afp	pore size	connectivity	0.0432
afp	node density	fiber length	0.0416
afp	node density	connectivity	0.0156
afp	node density	fiber OI	0.0089
afp	fiber diameter	connectivity	0.0418
afp	porosity	fiber length	0.0134
afp	porosity	connectivity	0.0349
afp	porosity	fiber OI	0.001
hnf4	pore size	connectivity	0.042
hnf4	pore size	fiber OI	0.0097
hnf4	porosity	connectivity	0.005
hnf4	connectivity	fiber OI	0.0085
cxcr4	porosity	connectivity	0.0471
foxa2	node density	fiber length	0.0279
foxa2	node density	fiber OI	0.0166
foxa2	porosity	fiber length	0.0387
foxa2	porosity	fiber OI	0.0231
ttr	pore size	fiber diameter	0.0442
ttr	node density	connectivity	0.0309
gata4	pore size	connectivity	0.0034

Table A2 (continued

gata4	node density	fiber length	0.0336
gata4	node density	connectivity	0.003
gata4	node density	fiber OI	0.0179
gata4	fiber diameter	connectivity	0.0091
gata4	porosity	fiber length	0.0082
gata4	porosity	connectivity	0.0475
gata4	porosity	fiber OI	0.0089

Table A.3. Parameters in the reduced G1 ODE model

Parameter	Value	Description
sD	3 mM/hr	Basal synthesis rates
sI	3 mM/hr	
sE	0 mM/hr	
dD	3.2 hr^{-1}	Degradation rates
dI	0.6 hr^{-1}	
dE	2.25 hr^{-1}	
d(D.I)	3 hr^{-1}	
d(E.I)	3 hr^{-1}	
k1	240 $mM^{-1}hr^{-1}$	D-I Dimerization constant
k2	12 hr^{-1}	D-I Dissociation constant
k3	1 $mM^{-1}hr^{-1}$	Rate constant for phosphorylation of I by E
k4	120 $mM^{-1}hr^{-1}$	E-I Dimerization constant
k5	24 hr^{-1}	I-I Dissociation constant
k6	10 mM/hr	Kinetic parameters for effect of cyclin E and D on cyclin E
k7	1.5 mM/hr	
k8	1.01 mM^2	
k9	1.01 mM	
H	2	Hill number
D(0)	0 mM	Initial conditions
I(0)	0.3 mM	
E(0)	0 mM	
D.I(0)	0 mM	
E.I(0)	0 mM	

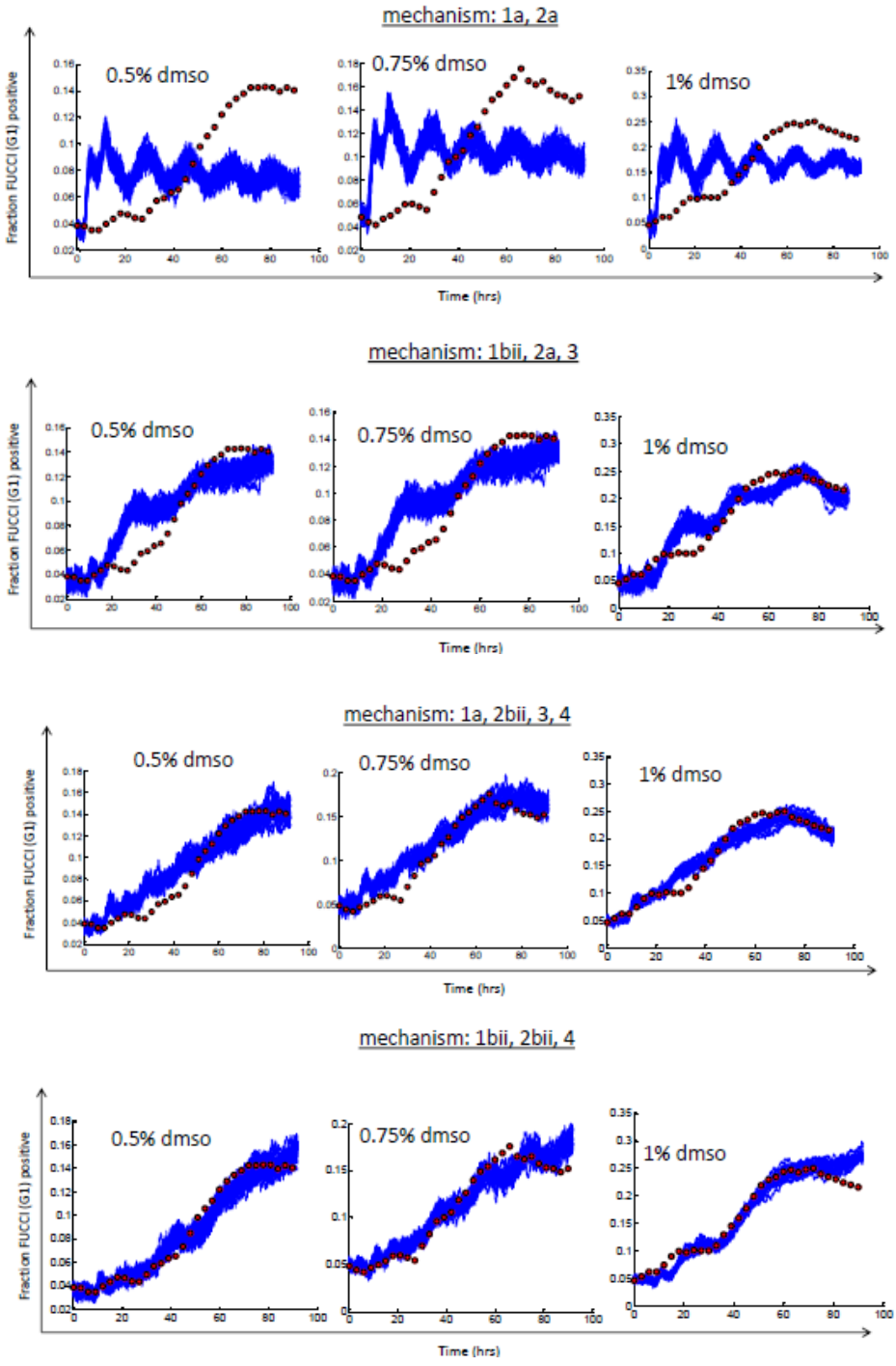

Figure A.1. Dynamics of ensemble model resulting from different mechanism alternatives

For each alternative, parameters were optimized for each of the three DMSO concentrations. Blue band: 100 stochastic simulations. Mechanism listed above each set of dynamics (see 4.3.3 for descriptions of mechanisms)

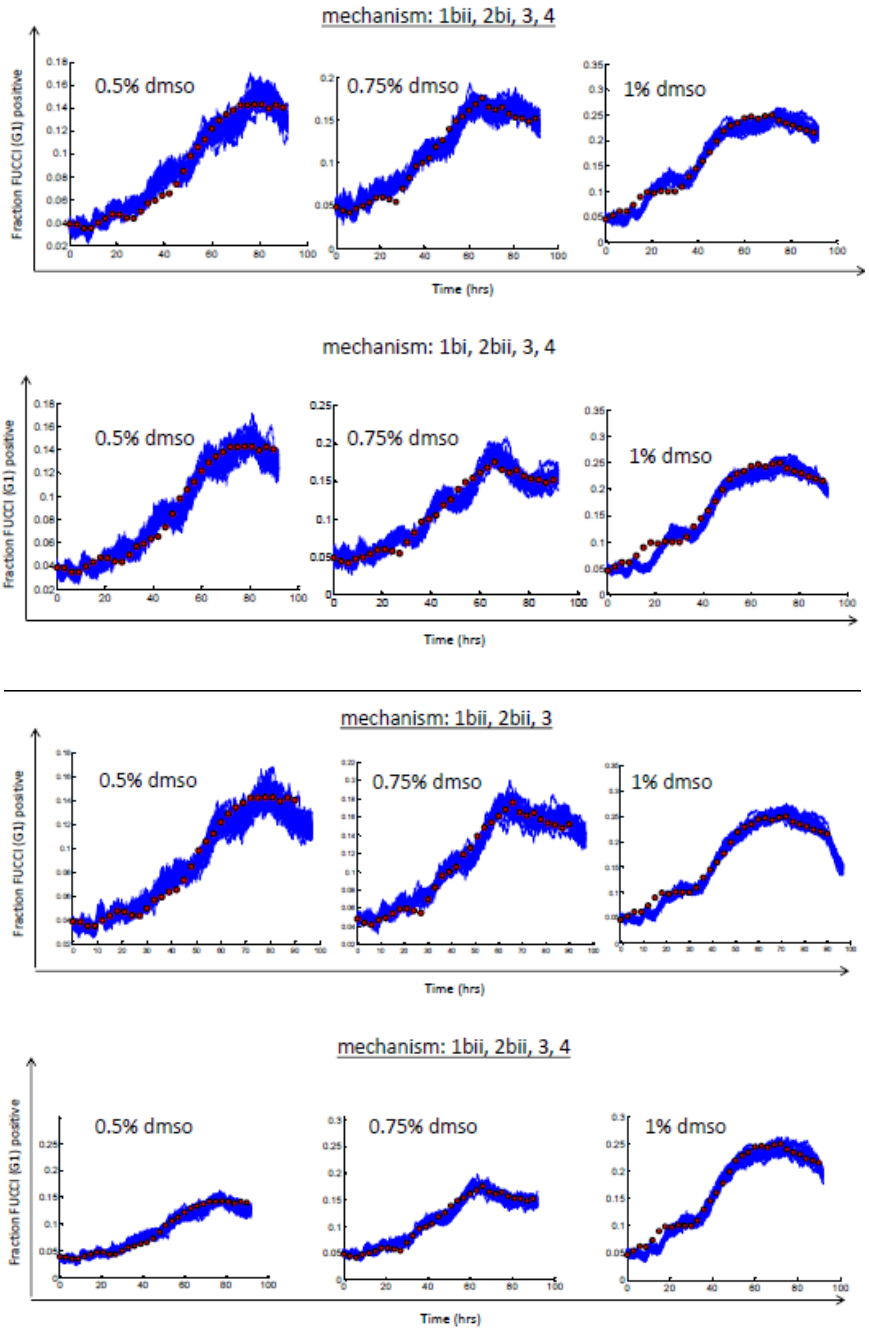

Figure A.2. Dynamics of ensemble model resulting from further mechanism alternatives

For each alternative, parameters were optimized for each of the three DMSO concentrations. Blue band: 100 stochastic simulations. Mechanism listed above each set of dynamics (see 4.3.3 for descriptions of mechanisms)

Table A.4. Variables used in ensemble model, Equation 4.8

Variable	Description
τ_{G1}	G1 residence time
α	Maximum G1 residence time of cells
$\tau_{G1,0}$	G1 residence time of undifferentiated cells. This parameter is normally distributed, with mean $\mu_{G1,undiff}$ and standard deviation $\sigma_{G1,undiff}$
$\mu_{G1,undiff}$	Mean of $\tau_{G1,0}$, from synchronization experiments
$\sigma_{G1,undiff}$	Standard deviation of $\tau_{G1,0}$, from synchronization experiments
γ	Exponential parameter governing the rate at which G1 will lengthen with time. This parameter is normally distributed, with mean μ_γ and standard deviation σ_γ
μ_γ	Mean of γ
σ_γ	Standard deviation of γ
T^*	Time at which cells are primed for differentiation. If a cell in the G1 phase is probabilistically chosen to differentiate, the cell is primed for differentiation and G1 lengthening upon the next cell cycle, and T^* is the time at which it exists the phase.
tstop	Time at which further differentiation ceases: G1 time stops lengthening and no further cells are primed for differentiation

Table A.5. Primer sequences used during PCR analysis of differentiation of hESC

Gene	Left sequence	Right sequence
Gapdh	acg acc act ttg tca agc tca ttt	gca gtg agg gtc tct ctc ttc ctc
Oct4	ctg ggt tga tcc tcg gac ct	cac aga act cat acg gcg gg
BrachT	tgc ttc cct gag acc cag tt	gat cac ttc ttt cct ttg cat caa g

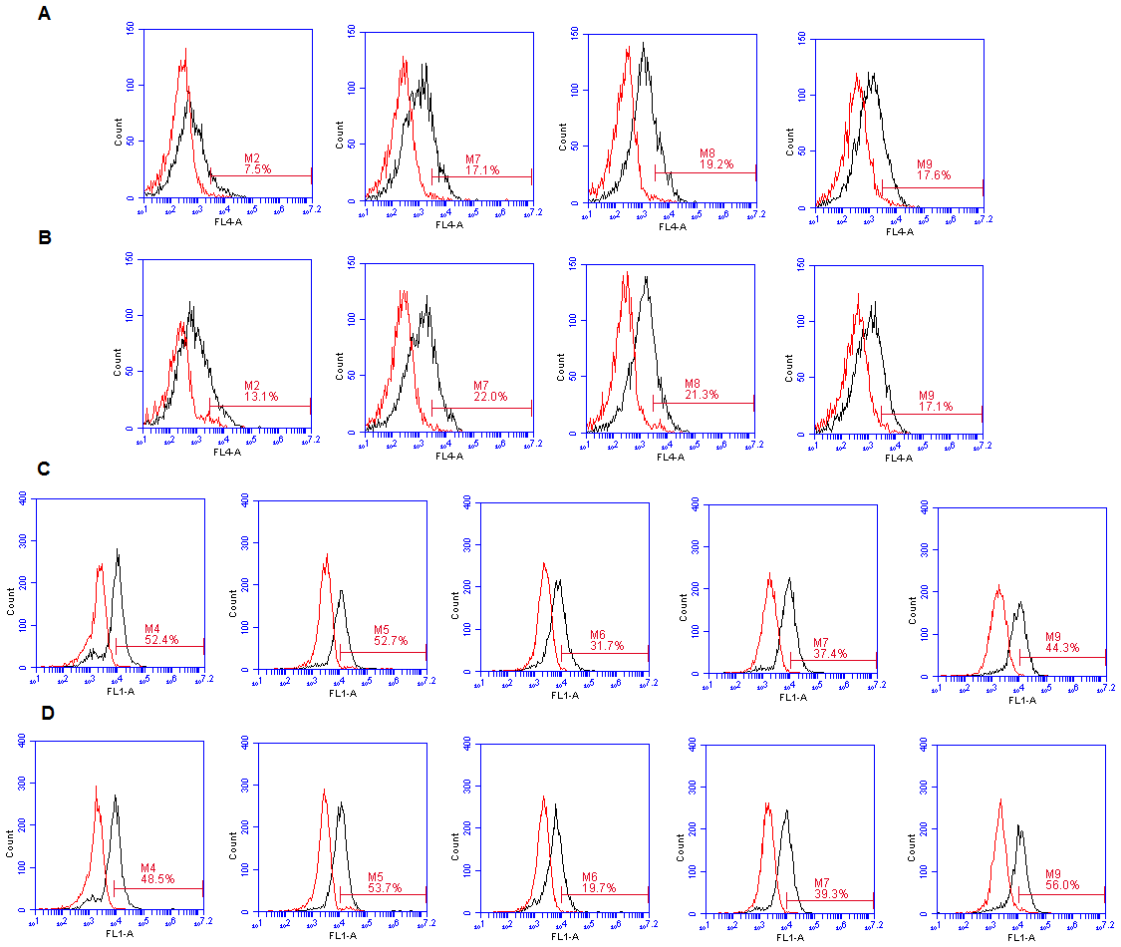

Figure A.3. Flow cytometry of cells positive for specific markers

Red histogram: negative (secondary antibody only) sample. Black histogram: stained sample. Red bar is gated beyond the negative control to denote the positive sample population. (a,b): Sox17 analysis for Conditions A and B, respectively. From left to right: Day 1-4. (c,d): CXCR4 analysis for Conditions A and B, respectively. From left to right: Day 1-5.

Table A.6. Definitions of the parameters used in the population based model

Parameter	Definition
a0max	maximum value of 'a' in initial cell population, first differentiation stage
a0max2	maximum value of 'a' in initial cell population, second differentiation stage
a0min	minimum value of 'a' in initial cell population, first differentiation stage
a0min2	minimum value of 'a' in initial cell population, second differentiation stage
aa	used when determining the probability of a cell transferring to the alpha regime
amin	threshold on 'a' below which a cell is able to proliferate
aw	used when determining the probability of a cell transferring to the omega regime
ba	used when determining the probability of a cell transferring to the alpha regime
bprog	used in updating propensity in omega regime
bw	used when determining the probability of a cell transferring to the omega regime
ca	used when determining the probability of a cell transferring to the alpha regime
cw	used when determining the probability of a cell transferring to the omega regime
d	factor by which 'a' decreases in omega regime
da	used when determining the probability of a cell transferring to the alpha regime
dw	used when determining the probability of a cell transferring to the omega regime
lmax	maximum bound on cell life span
lmin	minimum bound on cell life span
nprog1[a]	used in determining magnitude of propensity update in omega regime for lineage 1
nprog2[a]	used in determining magnitude of propensity update in omega regime for lineage 2
nprog3[a]	used in determining magnitude of propensity update in omega regime for lineage 3
nprog4[a]	used in determining magnitude of propensity update in omega regime for lineage 4
nreg	used in updating propensity in alpha regime

Table A6. (continued)	
tDstop	time beyond which a cell enters into a senescent stage and will not die (counted from start of cell's life)
tg1	time a cell stays in the g1 phase of the cell cycle
tpmax	upper bound of time beyond which a cell enters into a senescent stage and will not proliferate (counted from start of cell's life)
tpmin	lower bound of time beyond which a cell enters into a senescent stage and will not proliferate (counted from start of cell's life)
xcom	threshold level of propensity beyond which a cell is considered committed, first differentiation stage
xcom2	threshold level of propensity beyond which a cell is considered committed, second differentiation stage

Only the sensitive parameters are shown. A '2' after the parameter denotes the parameter for the second stage of differentiation (mesendoderm to mesoderm and endoderm) as opposed to the first stage (hESC to mesendoderm and visceral endoderm). ME: mesendoderm, VE: visceral endoderm

BIBLIOGRAPHY

[1] Reubinoff B.E., Pera M.F., Fong C.-Y., Trounson A., Bongso A. 2000 Embryonic stem cell lines from human blastocysts: somatic differentiation in vitro. Nat Biotech. 18, 399-404.

[2] Thomson J.A., Itskovitz-Eldor J., Shapiro S.S., Waknitz M.A., Swiergiel J.J., Marshall V.S., Jones J.M. 1998 Embryonic Stem Cell Lines Derived from Human Blastocysts. Science. 282, 1145-7. (10.1126/science.282.5391.1145)

[3] Czechanski A., Byers C., Greenstein I., Schrode N., Donahue L.R., Hadjantonakis A.-K., Reinholdt L.G. 2014 Derivation and characterization of mouse embryonic stem cells from permissive and nonpermissive strains. Nat Protocols. 9, 559-74. (10.1038/nprot.2014.030)

[4] Fluckiger A.-C., Marcy G., Marchand M., Négre D., Cosset F.-L., Mitalipov S., Wolf D., Savatier P., Dehay C. 2006 Cell Cycle Features of Primate Embryonic Stem Cells. STEM CELLS. 24, 547-56. (10.1634/stemcells.2005-0194)

[5] Smith A.G. 2001 EMBRYO-DERIVED STEM CELLS: Of Mice and Men. Annual Review of Cell and Developmental Biology. 17, 435-62. (doi:10.1146/annurev.cellbio.17.1.435)

[6] Murry C.E., Keller G. 2008 Differentiation of Embryonic Stem Cells to Clinically Relevant Populations: Lessons from Embryonic Development. Cell. 132, 661-80. (http://dx.doi.org/10.1016/j.cell.2008.02.008)

[7] Centers for Disease Control and Prevention. National Diabetes Statistics Report: Estimates of Diabetes and Its Burden in the United States, 2014. Atlanta, GA: US Department of Health and Human Services; 2014.

[8] Kahn C., Joslin E. Joslin's diabetes mellitus. Philadelphia: Lippincott Williams & Willkins; 2005.

[9] Silverthorn D. Human physiology: an integrated approach. Upper Saddle River, NJ: Prentice Hall; 1997.

[10] Hirshberg B., Rother K.I., Digon B.J., Lee J., Gaglia J.L., Hines K., Read E.J., Chang R., Wood B.J., Harlan D.M. 2003 Benefits and Risks of Solitary Islet Transplantation for Type 1 Diabetes Using Steroid-Sparing Immunosuppression: The National Institutes of Health experience. Diabetes Care. 26, 3288-95. (10.2337/diacare.26.12.3288)

[11] Shapiro A.M.J., Lakey J.R.T., Ryan E.A., Korbutt G.S., Toth E., Warnock G.L., Kneteman N.M., Rajotte R.V. 2000 Islet Transplantation in Seven Patients with Type 1 Diabetes Mellitus Using a Glucocorticoid-Free Immunosuppressive Regimen. New England Journal of Medicine. 343, 230-8. (doi:10.1056/NEJM200007273430401)

[12] Fu X., Xu Y. 2011 Self-renewal and scalability of human embryonic stem cells for human therapy. Regenerative Medicine. 6, 327-34. (10.2217/rme.11.18)

[13] Xu C., Inokuma M.S., Denham J., Golds K., Kundu P., Gold J.D., Carpenter M.K. 2001 Feeder-free growth of undifferentiated human embryonic stem cells. Nat Biotech. 19, 971-4.

[14] Bain G., Kitchens D., Yao M., Huettner J.E., Gottlieb D.I. 1995 Embryonic Stem Cells Express Neuronal Properties in Vitro. Developmental Biology. 168, 342-57. (10.1006/dbio.1995.1085)

[15] D'Amour K.A., Agulnick A.D., Eliazer S., Kelly O.G., Kroon E., Baetge E.E. 2005 Efficient differentiation of human embryonic stem cells to definitive endoderm. Nat Biotech. 23, 1534-41. (http://www.nature.com/nbt/journal/v23/n12/suppinfo/nbt1163_S1.html)

[16] Johansson B.M., Wiles M.V. 1995 Evidence for involvement of activin A and bone morphogenetic protein 4 in mammalian mesoderm and hematopoietic development. Molecular and Cellular Biology. 15, 141-51.

[17] D'Amour K.A., Bang A.G., Eliazer S., Kelly O.G., Agulnick A.D., Smart N.G., Moorman M.A., Kroon E., Carpenter M.K., Baetge E.E. 2006 Production of pancreatic hormone-expressing endocrine cells from human embryonic stem cells. Nat Biotech. 24, 1392-401. (http://www.nature.com/nbt/journal/v24/n11/suppinfo/nbt1259_S1.html)

[18] Kim S.-H., Turnbull J., Guimond S. 2011 Extracellular matrix and cell signalling: the dynamic cooperation of integrin, proteoglycan and growth factor receptor. Journal of Endocrinology. 209, 139-51. (10.1530/joe-10-0377)

[19] Sun Y., Chen C.S., Fu J. 2012 Forcing Stem Cells to Behave: A Biophysical Perspective of the Cellular Microenvironment. Annual Review of Biophysics. 41, 519-42. (doi:10.1146/annurev-biophys-042910-155306)

[20] Engler A.J., Sen S., Sweeney H.L., Discher D.E. 2006 Matrix Elasticity Directs Stem Cell Lineage Specification. Cell. 126, 677-89. (10.1016/j.cell.2006.06.044)

[21] Chambers I., Colby D., Robertson M., Nichols J., Lee S., Tweedie S., Smith A. 2003 Functional Expression Cloning of Nanog, a Pluripotency Sustaining Factor in Embryonic Stem Cells. Cell. 113, 643-55. (http://dx.doi.org/10.1016/S0092-8674(03)00392-1)

[22] Mitsui K., Tokuzawa Y., Itoh H., Segawa K., Murakami M., Takahashi K., Maruyama M., Maeda M., Yamanaka S. 2003 The Homeoprotein Nanog Is Required for Maintenance of Pluripotency in Mouse Epiblast and ES Cells. Cell. 113, 631-42. (http://dx.doi.org/10.1016/S0092-8674(03)00393-3)

[23] Boyer L.A., Lee T.I., Cole M.F., Johnstone S.E., Levine S.S., Zucker J.P., Guenther M.G., Kumar R.M., Murray H.L., Jenner R.G., et al. 2005 Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells. Cell. 122, 947-56. (http://dx.doi.org/10.1016/j.cell.2005.08.020)

[24] Loh Y.-H., Wu Q., Chew J.-L., Vega V.B., Zhang W., Chen X., Bourque G., George J., Leong B., Liu J., et al. 2006 The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nat Genet. 38, 431-40. (http://www.nature.com/ng/journal/v38/n4/suppinfo/ng1760_S1.html)

[25] Chickarmane V., Troein C., Nuber U.A., Sauro H.M., Peterson C. 2006 Transcriptional Dynamics of the Embryonic Stem Cell Switch. PLoS Comput Biol. 2, e123. (10.1371/journal.pcbi.0020123)

[26] Chickarmane V., Peterson C. 2008 A Computational Model for Understanding Stem Cell, Trophectoderm and Endoderm Lineage Determination. PLoS ONE. 3, e3478. (10.1371/journal.pone.0003478)

[27] MacArthur B.D., Please C.P., Oreffo R.O.C. 2008 Stochasticity and the Molecular Mechanisms of Induced Pluripotency. PLoS ONE. 3, e3086. (10.1371/journal.pone.0003086)

[28] Glauche I., Herberg M., Roeder I. 2010 Nanog Variability and Pluripotency Regulation of Embryonic Stem Cells - Insights from a Mathematical Model Analysis. PLoS ONE. 5, e11238. (10.1371/journal.pone.0011238)

[29] Prudhomme W., Daley G.Q., Zandstra P., Lauffenburger D.A. 2004 Multivariate proteomic analysis of murine embryonic stem cell self-renewal versus differentiation signaling. Proceedings of the National Academy of Sciences of the United States of America. 101, 2900-5. (10.1073/pnas.0308768101)

[30] Woolf P.J., Prudhomme W., Daheron L., Daley G.Q., Lauffenburger D.A. 2005 Bayesian analysis of signaling networks governing embryonic stem cell fate decisions. Bioinformatics. 21, 741-53. (10.1093/bioinformatics/bti056)

[31] Mahdavi A., Davey R.E., Bhola P., Yin T., Zandstra P.W. 2007 Sensitivity Analysis of Intracellular Signaling Pathway Kinetics Predicts Targets for Stem Cell Fate Control. PLoS Comput Biol. 3, e130. (10.1371/journal.pcbi.0030130)

[32] Viswanathan S., Benatar T., Rose-John S., Lauffenburger D.A., Zandstra P.W. 2002 Ligand/Receptor Signaling Threshold (LIST) Model Accounts for gp130-Mediated Embryonic Stem Cell Self-Renewal Responses to LIF and HIL-6. STEM CELLS. 20, 119-38. (10.1634/stemcells.20-2-119)

[33] Viswanathan S., Zandstra P. 2003 Towards predictive models of stem cell fate. Cytotechnology. 41, 75-92. (10.1023/a:1024866504538)

[34] Prudhomme W.A., Duggar K.H., Lauffenburger D.A. 2004 Cell population dynamics model for deconvolution of murine embryonic stem cell self-renewal and differentiation responses to cytokines and extracellular matrix. Biotechnology and Bioengineering. 88, 264-72. (10.1002/bit.20244)

[35] Pisu M., Concas A., Fadda S., Cincotti A., Cao G. 2008 A simulation model for stem cells differentiation into specialized cells of non-connective tissues. Computational Biology and Chemistry. 32, 338-44. (http://dx.doi.org/10.1016/j.compbiolchem.2008.06.001)

[36] Henson M.A. 2003 Dynamic modeling of microbial cell populations. Current Opinion in Biotechnology. 14, 460-7. (http://dx.doi.org/10.1016/S0958-1669(03)00104-6)

[37] Glauche I., Cross M., Loeffler M., Roeder I. 2007 Lineage Specification of Hematopoietic Stem Cells: Mathematical Modeling and Biological Implications. STEM CELLS. 25, 1791-9. (10.1634/stemcells.2007-0025)

[38] Becker K.A., Ghule P.N., Therrien J.A., Lian J.B., Stein J.L., van Wijnen A.J., Stein G.S. 2006 Self-renewal of human embryonic stem cells is supported by a shortened G1 cell cycle phase. Journal of Cellular Physiology. 209, 883-93. (10.1002/jcp.20776)

[39] White J., Dalton S. 2005 Cell cycle control of embryonic stem cells. Stem Cell Reviews and Reports. 1, 131-8. (10.1385/scr:1:2:131)

[40] Becker K.A., Ghule P.N., Lian J.B., Stein J.L., van Wijnen A.J., Stein G.S. 2010 Cyclin D2 and the CDK substrate p220NPAT are required for self-renewal of human embryonic stem cells. Journal of Cellular Physiology. 222, 456-64. (10.1002/jcp.21967)

[41] Calder A., Ivana Roth-Albin, Sonam Bhatia, Carlos Pilquil, Jong Hee Lee, Mick Bhatia, Marilyne Levadoux-Martin, Jamie McNicol, Jennifer Russell, Tony Collins, and Jonathan S. Draper. 2013 Lengthened G1 Phase Indicates Differentiation Status in Human Embryonic Stem Cells Stem Cells and Development. 22, 17.

[42] Gérard C., Goldbeter A. 2009 Temporal self-organization of the cyclin/Cdk network driving the mammalian cell cycle. Proceedings of the National Academy of Sciences. 106, 21643-8. (10.1073/pnas.0903827106)

[43] Toettcher J.E., Loewer A., Ostheimer G.J., Yaffe M.B., Tidor B., Lahav G. 2009 Distinct mechanisms act in concert to mediate cell cycle arrest. Proceedings of the National Academy of Sciences. 106, 785-90. (10.1073/pnas.0806196106)

[44] Iwamoto K., Tashima Y., Hamada H., Eguchi Y., Okamoto M. 2008 Mathematical modeling and sensitivity analysis of G1/S phase in the cell cycle including the DNA-damage signal transduction pathway. Biosystems. 94, 109-17. (http://dx.doi.org/10.1016/j.biosystems.2008.05.016)

[45] Pfeuty B. 2012 Dynamical principles of cell-cycle arrest: Reversible, irreversible, and mixed strategies. Physical Review E. 86, 021917.

[46] Pfeuty B. 2012 Strategic Cell-Cycle Regulatory Features That Provide Mammalian Cells with Tunable G1 Length and Reversible G1 Arrest. PLoS ONE. 7, e35291. (10.1371/journal.pone.0035291)

[47] Aguda B.D., Kim Y., Piper-Hunter M.G., Friedman A., Marsh C.B. 2008 MicroRNA regulation of a cancer network: Consequences of the feedback loops involving miR-17-92, E2F, and Myc. Proceedings of the National Academy of Sciences. 105, 19678-83. (10.1073/pnas.0811166106)

[48] Novák B., Tyson J.J. 2004 A model for restriction point control of the mammalian cell cycle. Journal of Theoretical Biology. 230, 563-79. (10.1016/j.jtbi.2004.04.039)

[49] Conradie R., Bruggeman F.J., Ciliberto A., Csikász-Nagy A., Novák B., Westerhoff H.V., Snoep J.L. 2010 Restriction point control of the mammalian cell cycle via the cyclin E/Cdk2:p27 complex. FEBS Journal. 277, 357-67. (10.1111/j.1742-4658.2009.07473.x)

[50] Swat M., Kel A., Herzel H. 2004 Bifurcation analysis of the regulatory modules of the mammalian G1/S transition. Bioinformatics. 20, 1506-11. (10.1093/bioinformatics/bth110)

[51] Yao G., Lee T.J., Mori S., Nevins J.R., You L. 2008 A bistable Rb-E2F switch underlies the restriction point. Nat Cell Biol. 10, 476-82. (http://www.nature.com/ncb/journal/v10/n4/suppinfo/ncb1711_S1.html)

[52] Liu Y.-H., Bi J.-X., Zeng A.-P., Yuan J.-Q. 2007 A Population Balance Model Describing the Cell Cycle Dynamics of Myeloma Cell Cultivation. Biotechnology Progress. 23, 1198-209. (10.1021/bp070152z)

[53] Stukalin E.B., Aifuwa I., Kim J.S., Wirtz D., Sun S.X. 2013 Age-dependent stochastic models for understanding population fluctuations in continuously cultured cells. Journal of The Royal Society Interface. 10, (10.1098/rsif.2013.0325)

[54] Altinok A., Lévi F., Goldbeter A. 2007 A cell cycle automaton model for probing circadian patterns of anticancer drug delivery. Advanced Drug Delivery Reviews. 59, 1036-53. (http://dx.doi.org/10.1016/j.addr.2006.09.022)

[55] Altinok A., Lévi F., Goldbeter A. 2009 Identifying mechanisms of chronotolerance and chronoefficacy for the anticancer drugs 5-fluorouracil and oxaliplatin by computational modeling. European Journal of Pharmaceutical Sciences. 36, 20-38. (http://dx.doi.org/10.1016/j.ejps.2008.10.024)

[56] Gurkan E., Schupp J.E., Aziz M.A., Kinsella T.J., Loparo K.A. 2007 Probabilistic Modeling of DNA Mismatch Repair Effects on Cell Cycle Dynamics and Iododeoxyuridine-DNA Incorporation. Cancer Research. 67, 10993-1000. (10.1158/0008-5472.can-07-0966)

[57] Hillen T., de Vries G., Gong J., Finlay C. 2010 From cell population models to tumor control probability: Including cell cycle effects. Acta Oncologica. 49, 1315-23. (doi:10.3109/02841861003631487)

[58] Basse B., Ubezio P. 2007 A Generalised Age- and Phase-Structured Model of Human Tumour Cell Populations Both Unperturbed and Exposed to a Range of Cancer Therapies. Bull Math Biol. 69, 1673-90. (10.1007/s11538-006-9185-6)

[59] Priori L., Ubezio P. 1996 Mathematical modelling and computer simulation of cell synchrony. Methods Cell Sci. 18, 83-91. (10.1007/bf00122158)

[60] Orlando D., Lin C.Y., Bernard A., Iversen E.S., Hartemink A.J., Haase S.B. 2007 A Probabilistic Model for Cell Cycle Distributions in Synchrony Experiments. Cell Cycle. 6, 478-88.

[61] Billy F., Clairambaultt J., Fercoq O., Gaubertt S., Lepoutre T., Ouillon T., Saito S. 2014 Synchronisation and control of proliferation in cycling cell population models with age structure. Mathematics and Computers in Simulation. 96, 66-94. (http://dx.doi.org/10.1016/j.matcom.2012.03.005)

[62] Chemmangattuvalappil N., Task K., Banerjee I. 2012 An integer optimization algorithm for robust identification of non-linear gene regulatory networks. BMC Systems Biology. 6, 14.

[63] Stelling J., Sauer U., Szallasi Z., Doyle Iii F.J., Doyle J. 2004 Robustness of Cellular Functions. Cell. 118, 675-85. (http://dx.doi.org/10.1016/j.cell.2004.09.008)

[64] Whitacre J.M. 2012 Biological Robustness: Paradigms, Mechanisms, and Systems Principles. Frontiers in Genetics. 3, (10.3389/fgene.2012.00067)

[65] Wang Z., Zhang J. 2009 Abundant Indispensable Redundancies in Cellular Metabolic Networks. Genome Biology and Evolution. 1, 11.
[66] Ma H.-W., Zeng A.-P. 2003 The connectivity structure, giant strong component and centrality of metabolic networks. Bioinformatics. 19, 1423-30. (10.1093/bioinformatics/btg177)

[67] Leclerc R.D. 2008 Survival of the sparsest: robust gene networks are parsimonious. Mol Syst Biol. 4, (http://www.nature.com/msb/journal/v4/n1/suppinfo/msb200852_S1.html)

[68] Shen-Orr S.S., Milo R., Mangan S., Alon U. 2002 Network motifs in the transcriptional regulation network of Escherichia coli. Nat Genet. 31, 64-8. (http://www.nature.com/ng/journal/v31/n1/suppinfo/ng881_S1.html)

[69] Costanzo M.C., Crawford M.E., Hirschman J.E., Kranz J.E., Olsen P., Robertson L.S., Skrzypek M.S., Braun B.R., Hopkins K.L., Kondu P., et al. 2001 YPD™, PombePD™ and

WormPD™: model organism volumes of the BioKnowledge™ Library, an integrated resource for protein information. Nucleic Acids Research. 29, 75-9. (10.1093/nar/29.1.75)

[70] Davidson E.H., Rast J.P., Oliveri P., Ransick A., Calestani C., Yuh C.-H., Minokawa T., Amore G., Hinman V., Arenas-Mena C., et al. 2002 A Genomic Regulatory Network for Development. Science. 295, 1669-78. (10.1126/science.1069883)

[71] Banerjee I., Maiti S., Parashurama N., Yarmush M. 2010 An integer programming formulation to identify the sparse network architecture governing differentiation of embryonic stem cells. Bioinformatics. 26, 1332-9. (10.1093/bioinformatics/btq139)

[72] Thomas R., Mehrotra S., Papoutsakis E.T., Hatzimanikatis V. 2004 A model-based optimization framework for the inference on gene regulatory networks from DNA array data. Bioinformatics. 20, 3221-35. (10.1093/bioinformatics/bth389)

[73] Kikuchi S., Tominaga D., Arita M., Takahashi K., Tomita M. 2003 Dynamic modeling of genetic networks using genetic algorithm and S-system. Bioinformatics. 19, 643-50. (10.1093/bioinformatics/btg027)

[74] Papadimitriou C.H., Steiglitz K. Combinatorial optimization: Algorithms and complexity. Mineola, NY: Dover; 1998.

[75] Yao X. Evolutionary computation: Theory and applications. Singapore: World Scientific Publishing; 1999.

[76] Gutenkunst R.N., Waterfall J.J., Casey F.P., Brown K.S., Myers C.R., Sethna J.P. 2007 Universally Sloppy Parameter Sensitivities in Systems Biology Models. PLoS Comput Biol. 3, e189. (10.1371/journal.pcbi.0030189)

[77] Efron B., Tibshirani R.J. An introduction to bootstrap. New York: Chapman and Hall; 1993.

[78] Wehrens R., Putter H., Buydens L.M.C. 2000 The bootstrap: a tutorial. Chemometrics and Intelligent Laboratory Systems. 54, 18.

[79] Vilela M., Chou I.-C., Vinga S., Vasconcelos A., Voit E., Almeida J. 2008 Parameter optimization in S-system models. BMC Systems Biology. 2, 35.

[80] Liu P.-K., Wang F.-S. 2008 Inference of biochemical network models in S-system using multiobjective optimization approach. Bioinformatics. 24, 1085-92. (10.1093/bioinformatics/btn075)

[81] Sutton M.D., Smith B.T., Godoy V.G., Walker G.C. 2000 THE SOS RESPONSE: Recent Insights into umuDC-Dependent Mutagenesis and DNA Damage Tolerance. Annual Review of Genetics. 34, 479-97. (doi:10.1146/annurev.genet.34.1.479)

[82] Ronen M., Rosenberg R., Shraiman B.I., Alon U. 2002 Assigning numbers to the arrows: Parameterizing a gene regulation network by using accurate expression kinetics. Proceedings of the National Academy of Sciences. 99, 10555-60. (10.1073/pnas.152046799)

[83] Index of /mcb/UriAlon/Papers/SOSData. [http://www.weizmann.ac.il/mcb/UriAlon/Papers/SOSData/].

[84] Kimura S., Nakayama S., Hatakeyama M. 2009 Genetic network inference as a series of discrimination tasks. Bioinformatics. 25, 918-25. (10.1093/bioinformatics/btp072)

[85] Task K., D'Amore A., Singh S., Candiello J., Jaramillo M., Wagner W.R., Kumta P., Banerjee I. 2014 Systems level approach reveals the correlation of endoderm differentiation of mouse embryonic stem cells with specific microstructural cues of fibrin gels. Journal of The Royal Society Interface. 11, (10.1098/rsif.2014.0009)

[86] Yasuda H., Kuroda S., Shichinohe H., Kamei S., Kawamura R., Iwasaki Y. 2010 Effect of biodegradable fibrin scaffold on survival, migration, and differentiation of transplanted bone marrow stromal cells after cortical injury in rats. Journal of Neurosurgery. 112, 336-44. (doi:10.3171/2009.2.JNS08495)

[87] D'Amore A., Stella J.A., Wagner W.R., Sacks M.S. 2010 Characterization of the complete fiber network topology of planar fibrous tissues and scaffolds. Biomaterials. 31, 5345-54. (10.1016/j.biomaterials.2010.03.052)

[88] Hutter J.L., Bechhoefer J. 1993 Calibration of atomic‐force microscope tips. Review of Scientific Instruments. 64, 1868-73. (10.1063/1.1143970)

[89] Radmacher M., Fritz M., Hansma P.K. 1995 Imaging soft samples with the atomic force microscope: gelatin in water and propanol. Biophysical Journal. 69, 264-70. (10.1016/s0006-3495(95)79897-6)

[90] Rhee S.K., Quist A.P., Lal R. 1998 Amyloid β Protein-(1–42) Forms Calcium-permeable, Zn2+-sensitive Channel. Journal of Biological Chemistry. 273, 13379-82. (10.1074/jbc.273.22.13379)

[91] Candiello J., Balasubramani M., Schreiber E.M., Cole G.J., Mayer U., Halfter W., Lin H. 2007 Biomechanical properties of native basement membranes. FEBS Journal. 274, 2897-908. (10.1111/j.1742-4658.2007.05823.x)

[92] Quist A.P., Rhee S.K., Lin H., Lal R. 2000 Physiological Role of Gap-Junctional Hemichannels. The Journal of Cell Biology. 148, 1063-74. (10.1083/jcb.148.5.1063)

[93] Sneddon I.N. 1965 The relation between load and penetration in the axisymmetric boussinesq problem for a punch of arbitrary profile. International Journal of Engineering Science. 3, 47-57. (10.1016/0020-7225(65)90019-4)

[94] Almqvist E.G., Becker C., Bondeson A.-G., Bondeson L., Svensson J. 2004 Early parathyroidectomy increases bone mineral density in patients with mild primary hyperparathyroidism: A prospective and randomized study. Surgery. 136, 1281-8. (10.1016/j.surg.2004.06.059)

[95] Amoroso N., D'Amore A., Hong Y., Rivera C.P., Sacks M.S., Wagner W.R. 2012 Microstructure Manipulation to Tune Bending Stiffness in Electrospun Scaffolds for Heart Valve Tissue Engineering. In press on Acta Biomaterialia, doi: 101016/jactbio201208002.

[96] Amoroso N.J., D'Amore A., Hong Y., Wagner W.R., Sacks M.S. 2011 Elastomeric Electrospun Polyurethane Scaffolds: The Interrelationship Between Fabrication Conditions, Fiber Topology, and Mechanical Properties. Advanced Materials. 23, 106-11. (10.1002/adma.201003210)

[97] Wolf M.T., Daly K.A., Brennan-Pierce E.P., Johnson S.A., Carruthers C.A., D'Amore A., Nagarkar S.P., Velankar S.S., Badylak S.F. 2012 A hydrogel derived from decellularized dermal extracellular matrix. Biomaterials. 33, 7028-38. (10.1016/j.biomaterials.2012.06.051)

[98] Yasunaga M., Tada S., Torikai-Nishikawa S., Nakano Y., Okada M., Jakt L.M., Nishikawa S., Chiba T., Era T., Nishikawa S.-I. 2005 Induction and monitoring of definitive and visceral endoderm differentiation of mouse ES cells. Nat Biotech. 23, 1542-50. (http://www.nature.com/nbt/journal/v23/n12/suppinfo/nbt1167_S1.html)

[99] Ying Q.-L., Stavridis M., Griffiths D., Li M., Smith A. 2003 Conversion of embryonic stem cells into neuroectodermal precursors in adherent monoculture. Nat Biotech. 21, 183-6. (http://www.nature.com/nbt/journal/v21/n2/suppinfo/nbt780_S1.html)

[100] Jaramillo M., Singh S.S., Velankar S., Kumta P.N., Banerjee I. 2012 Inducing endoderm differentiation by modulating mechanical properties of soft substrates. Journal of Tissue Engineering and Regenerative Medicine. n/a-n/a. (10.1002/term.1602)

[101] Teo A.K.K., Ali Y., Wong K.Y., Chipperfield H., Sadasivam A., Poobalan Y., Tan E.K., Wang S.T., Abraham S., Tsuneyoshi N., et al. 2012 Activin and BMP4 Synergistically Promote Formation of Definitive Endoderm in Human Embryonic Stem Cells. STEM CELLS. 30, 631-42. (10.1002/stem.1022)

[102] Torres J., Prieto J., Durupt F.C., Broad S., Watt F.M. 2012 Efficient Differentiation of Embryonic Stem Cells into Mesodermal Precursors by BMP, Retinoic Acid and Notch Signalling. PLoS ONE. 7, e36405. (10.1371/journal.pone.0036405)

[103] Bashur C.A., Dahlgren L.A., Goldstein A.S. 2006 Effect of fiber diameter and orientation on fibroblast morphology and proliferation on electrospun poly(d,l-lactic-co-glycolic acid) meshes. Biomaterials. 27, 5681-8. (10.1016/j.biomaterials.2006.07.005)

[104] Duong H., Wu B., Tawil B. 2009 Modulation of 3D Fibrin Matrix Stiffness by Intrinsic Fibrinogen–Thrombin Compositions and by Extrinsic Cellular Activity Tissue Engineering Part A. 15, 1865-76.

[105] Badami A.S., Kreke M.R., Thompson M.S., Riffle J.S., Goldstein A.S. 2006 Effect of fiber diameter on spreading, proliferation, and differentiation of osteoblastic cells on electrospun poly(lactic acid) substrates. Biomaterials. 27, 596-606. (10.1016/j.biomaterials.2005.05.084)

[106] Ventura Ferreira M.S., Jahnen-Dechent W., Labude N., Bovi M., Hieronymus T., Zenke M., Schneider R.K., Neurs S. 2012 Cord blood-hematopoietic stem cell expansion in 3D fibrin scaffolds with stromal support. Biomaterials. 33, 6987-97. (http://dx.doi.org/10.1016/j.biomaterials.2012.06.029)

[107] Christman K.L., Vardanian A.J., Fang Q., Sievers R.E., Fok H.H., Lee R.J. 2004 Injectable Fibrin Scaffold Improves Cell Transplant Survival, Reduces Infarct Expansion, and Induces Neovasculature Formation in Ischemic Myocardium. Journal of the American College of Cardiology. 44, 654-60. (http://dx.doi.org/10.1016/j.jacc.2004.04.040)

[108] Bensaïd W., Triffitt J.T., Blanchat C., Oudina K., Sedel L., Petite H. 2003 A biodegradable fibrin scaffold for mesenchymal stem cell transplantation. Biomaterials. 24, 2497-502. (http://dx.doi.org/10.1016/S0142-9612(02)00618-X)

[109] Willerth S.M., Arendas K.J., Gottlieb D.I., Sakiyama-Elbert S.E. 2006 Optimization of fibrin scaffolds for differentiation of murine embryonic stem cells into neural lineage cells. Biomaterials. 27, 5990-6003. (10.1016/j.biomaterials.2006.07.036)

[110] Willerth S.M., Faxel T.E., Gottlieb D.I., Sakiyama-Elbert S.E. 2007 The Effects of Soluble Growth Factors on Embryonic Stem Cell Differentiation Inside of Fibrin Scaffolds. STEM CELLS. 25, 2235-44. (10.1634/stemcells.2007-0111)

[111] Vallée J.-P., Hauwel M., Lepetit-Coiffé M., Bei W., Montet-Abou K., Meda P., Gardier S., Zammaretti P., Kraehenbuehl T.P., Herrmann F., et al. 2012 Embryonic Stem Cell-Based Cardiopatches Improve Cardiac Function in Infarcted Rats. Stem Cells Translational Medicine. 1, 248-60. (10.5966/sctm.2011-0028)

[112] Soon A.S.C., Stabenfeldt S.E., Brown W.E., Barker T.H. 2010 Engineering fibrin matrices: The engagement of polymerization pockets through fibrin knob technology for the delivery and retention of therapeutic proteins. Biomaterials. 31, 1944-54. (http://dx.doi.org/10.1016/j.biomaterials.2009.10.060)

[113] Stabenfeldt S.E., Gourley M., Krishnan L., Hoying J.B., Barker T.H. 2012 Engineering fibrin polymers through engagement of alternative polymerization mechanisms. Biomaterials. 33, 535-44. (http://dx.doi.org/10.1016/j.biomaterials.2011.09.079)

[114] Perumcherry S.R., Chennazhi K.P., Nair S.V., Menon D., Afeesh R. 2011 A Novel Method for the Fabrication of Fibrin-Based Electrospun Nanofibrous Scaffold for Tissue-Engineering Applications. Tissue Engineering Part C: Methods. 17, 10.

[115] Herbert C.B., Nagaswami C., Bittner G.D., Hubbell J.A., Weisel J.W. 1998 Effects of fibrin micromorphology on neurite growth from dorsal root ganglia cultured in three-dimensional fibrin gels. Journal of Biomedical Materials Research. 40, 551-9. (10.1002/(sici)1097-4636(19980615)40:4<551::aid-jbm6>3.0.co;2-e)

[116] Karlon W.J., Covell J.W., McCulloch A.D., Hunter J.J., Omens J.H. 1998 Automated measurement of myofiber disarray in transgenic mice with ventricular expression of ras. The Anatomical Record. 252, 612-25. (10.1002/(sici)1097-0185(199812)252:4<612::aid-ar12>3.0.co;2-1)

[117] Chaudhuri B.B., Kundu P., Sarkar N. 1993 Detection and gradation of oriented texture. Pattern Recognition Letters. 14, 147-53. (http://dx.doi.org/10.1016/0167-8655(93)90088-U)

[118] Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. New York: Springer; 2001.

[119] Solana R.P., Chinchilli V.M., Carter W.H., Wilson J.D., Carchman R.A. 1987 The evaluation of biological interactions using response surface methodology. Cell Biology and Toxicology. 3, 263-77. (10.1007/bf00117864)

[120] Chen W.L.K., Likhitpanichkul M., Ho A., Simmons C.A. 2010 Integration of statistical modeling and high-content microscopy to systematically investigate cell–substrate interactions. Biomaterials. 31, 2489-97. (10.1016/j.biomaterials.2009.12.002)

[121] Nam J., Johnson J., Lannutti J.J., Agarwal S. 2011 Modulation of embryonic mesenchymal progenitor cell differentiation via control over pure mechanical modulus in electrospun nanofibers. Acta Biomaterialia. 7, 1516-24. (10.1016/j.actbio.2010.11.022)

[122] Evans N., Minelli C., Gentleman E., LaPointe V., Patankar S.N., Kallivretaki M., Chen X., Roberts C.J., Stevens M.M. 2009 Substrate Stiffness Affects Early Differentiation Events in Embryonic Stem Cells. European Cells and Materials. 18, 1-14.

[123] Gilbert P.M., Havenstrite K.L., Magnusson K.E.G., Sacco A., Leonardi N.A., Kraft P., Nguyen N.K., Thrun S., Lutolf M.P., Blau H.M. 2010 Substrate Elasticity Regulates Skeletal Muscle Stem Cell Self-Renewal in Culture. Science. 329, 1078-81. (10.1126/science.1191035)

[124] Huang X., Yang N., Fiore V.F., Barker T.H., Sun Y., Morris S.W., Ding Q., Thannickal V.J., Zhou Y. 2012 Matrix Stiffness–Induced Myofibroblast Differentiation Is Mediated by Intrinsic Mechanotransduction. American Journal of Respiratory Cell and Molecular Biology. 47, 340-8. (10.1165/rcmb.2012-0050OC)

[125] Li X., Huang Y., Zheng L., Liu H., Niu X., Huang J., Zhao F., Fan Y. 2013 Effect of substrate stiffness on the functions of rat bone marrow and adipose tissue derived mesenchymal stem cells in vitro. Journal of Biomedical Materials Research Part A. n/a-n/a. (10.1002/jbm.a.34774)

[126] Xue R., Li J.Y.-S., Yeh Y., Yang L., Chien S. 2013 Effects of matrix elasticity and cell density on human mesenchymal stem cells differentiation. Journal of Orthopaedic Research. 31, 1360-5. (10.1002/jor.22374)

[127] Candiello J., Singh S.S., Task K., Kumta P.N., Banerjee I. 2013 Early differentiation patterning of mouse embryonic stem cells in response to variations in alginate substrate stiffness. Journal of Biological Engineering. 7,

[128] Zhang X., Jaramillo M., Singh S., Kumta P., Banerjee I. 2012 Analysis of Regulatory Network Involved in Mechanical Induction of Embryonic Stem Cell Differentiation. PLoS ONE. 7, e35700. (10.1371/journal.pone.0035700)

[129] Hong H., Stegemann J.P. 2008 2D and 3D collagen and fibrin biopolymers promote specific ECM and integrin gene expression by vascular smooth muscle cells. Journal of Biomaterials Science, Polymer Edition. 19, 1279-93. (10.1163/156856208786052380)

[130] Liu C.Y., Nossel H.L., Kaplan K.L. 1979 The binding of thrombin by fibrin. Journal of Biological Chemistry. 254, 10421-5.

[131] Seegers W.H. 1947 Multiple Protein Interactions as Exhibited by the Blood-clotting Mechanism. The Journal of Physical and Colloid Chemistry. 51, 198-206. (10.1021/j150451a015)

[132] Mukhatyar V.J., Salmerón-Sánchez M., Rudra S., Mukhopadaya S., Barker T.H., García A.J., Bellamkonda R.V. 2011 Role of fibronectin in topographical guidance of neurite extension on electrospun fibers. Biomaterials. 32, 3958-68. (http://dx.doi.org/10.1016/j.biomaterials.2011.02.015)

[133] Dalby M.J., Gadegaard N., Tare R., Andar A., Riehle M.O., Herzyk P., Wilkinson C.D.W., Oreffo R.O.C. 2007 The control of human mesenchymal cell differentiation using nanoscale symmetry and disorder. Nat Mater. 6, 997-1003. (http://www.nature.com/nmat/journal/v6/n12/suppinfo/nmat2013_S1.html)

[134] Trappmann B., Gautrot J.E., Connelly J.T., Strange D.G.T., Li Y., Oyen M.L., Cohen Stuart M.A., Boehm H., Li B., Vogel V., et al. 2012 Extracellular-matrix tethering regulates stem-cell fate. Nat Mater. 11, 642-9. (http://www.nature.com/nmat/journal/v11/n7/abs/nmat3339.html#supplementary-information)

[135] Olivares-Navarrete R., Raz P., Zhao G., Chen J., Wieland M., Cochran D.L., Chaudhri R.A., Ornoy A., Boyan B.D., Schwartz Z. 2008 Integrin α2β1 plays a critical role in osteoblast response to micron-scale surface structure and surface energy of titanium substrates. Proceedings of the National Academy of Sciences. 105, 15767-72. (10.1073/pnas.0805420105)

[136] Weaver R.F. Molecular Biology. 3rd ed. New York: McGraw Hill; 2005.

[137] Blomen V.A., Boonstra J. 2007 Cell fate determination during G1 phase progression. Cell Mol Life Sci. 64, 3084-104. (10.1007/s00018-007-7271-z)

[138] Pauklin S., Vallier L. 2013 The Cell-Cycle State of Stem Cells Determines Cell Fate Propensity. Cell. 155, 135-47. (http://dx.doi.org/10.1016/j.cell.2013.08.031)

[139] Lange C., Calegari F. 2010 Cdks and cyclins link G₁ length and differentiation of embryonic, neural and hematopoietic stem cells. Cell Cycle. 9, 1893-900.
[140] Altinok A., Gonze D., Lévi F., Goldbeter A. 2011 An automaton model for the cell cycle. Interface Focus. 1, 36-47. (10.1098/rsfs.2010.0009)

[141] Smith J.A., Martin L. 1973 Do Cells Cycle? Proceedings of the National Academy of Sciences. 70, 1263-7.

[142] Sela Y., Molotski N., Golan S., Itskovitz-Eldor J., Soen Y. 2012 Human Embryonic Stem Cells Exhibit Increased Propensity to Differentiate During the G1 Phase Prior to Phosphorylation of Retinoblastoma Protein. STEM CELLS. 30, 1097-108. (10.1002/stem.1078)

[143] Roccio M., Schmitter D., Knobloch M., Okawa Y., Sage D., Lutolf M.P. 2013 Predicting stem cell fate changes by differential cell cycle progression patterns. Development. 140, 459-70. (10.1242/dev.086215)

[144] Singh Amar M., Chappell J., Trost R., Lin L., Wang T., Tang J., Wu H., Zhao S., Jin P., Dalton S. 2013 Cell-Cycle Control of Developmentally Regulated Transcription Factors Accounts for Heterogeneity in Human Pluripotent Cells. Stem Cell Reports. 1, 532-44. (10.1016/j.stemcr.2013.10.009)

[145] Chiorino G., Metz J.A.J., Tomasoni D., Ubezio P. 2001 Desynchronization Rate in Cell Populations: Mathematical Modeling and Experimental Data. Journal of Theoretical Biology. 208, 185-99. (http://dx.doi.org/10.1006/jtbi.2000.2213)

[146] Murphy J.S., D'Alisa R., Gershey E.L., Landsberger F.R. 1978 Kinetics of Desynchronization and Distribution of Generation Times in Synchronized Cell Population. Proceedings of the National Academy of Sciences of the United States of America. 75, 4.

[147] Ball D.A., Marchand J., Poulet M., Baumann W.T., Chen K.C., Tyson J.J., Peccoud J. 2011 Oscillatory Dynamics of Cell Cycle Proteins in Single Yeast Cells Analyzed by Imaging Cytometry. PLoS ONE. 6, e26272. (10.1371/journal.pone.0026272)

[148] Akaike H. 1974 A new look at the statistical model identification. Automatic Control, IEEE Transactions on. 19, 716-23. (10.1109/tac.1974.1100705)

[149] Johnson J.B., Omland K.S. Model selection in ecology and evolution. Trends in Ecology & Evolution. 19, 101-8. (10.1016/j.tree.2003.10.013)

[150] Becker K.A., Stein J.L., Lian J.B., van Wijnen A.J., Stein G.S. 2010 Human embryonic stem cells are pre-mitotically committed to self-renewal and acquire a lengthened G1 phase upon lineage programming. Journal of Cellular Physiology. 222, 103-10. (10.1002/jcp.21925)

[151] Golubev A. 2012 Transition probability in cell proliferation, stochasticity in cell differentiation, and the restriction point of the cell cycle in one package. Progress in Biophysics and Molecular Biology. 110, 87-96. (http://dx.doi.org/10.1016/j.pbiomolbio.2012.05.002)

[152] Jaramillo M., Mathew S., Task K., Barner S., Banerjee I. 2014 Potential for Pancreatic Maturation of Differentiating Human Embryonic Stem Cells Is Sensitive to the Specific Pathway of Definitive Endoderm Commitment. PLoS ONE. 9, e94307. (10.1371/journal.pone.0094307)

[153] Chetty S., Pagliuca F.W., Honore C., Kweudjeu A., Rezania A., Melton D.A. 2013 A simple tool to improve pluripotent stem cell differentiation. Nat Meth. 10, 553-6. (10.1038/nmeth.2442

http://www.nature.com/nmeth/journal/v10/n6/abs/nmeth.2442.html#supplementary-information)

[154] Neganova I., Zhang X., Atkinson S., Lako M. 2008 Expression and functional analysis of G1 to S regulatory components reveals an important role for CDK2 in cell cycle regulation in human embryonic stem cells. Oncogene. 28, 20-30. (http://www.nature.com/onc/journal/v28/n1/suppinfo/onc2008358s1.html)

[155] Egozi D., Shapira M., Paor G., Ben-Izhak O., Skorecki K., Hershko D.D. 2007 Regulation of the cell cycle inhibitor p27 and its ubiquitin ligase Skp2 in differentiation of human embryonic stem cells. The FASEB Journal. 21, 2807-17. (10.1096/fj.06-7758com)

[156] Miura T., Yongquan Luo, Irina Khrebtukova, Ralph Brandenberger, Daixing Zhou, R. Scott Thies, Tom Vasicek, Holly Young, Jane Lebkowski, Melissa K. Carpenter, and Dr. Mahendra S. Rao. 2004 Monitoring Early Differentiation Events in Human Embryonic Stem Cells by Massively Parallel Signature Sequencing and Expressed Sequence Tag Scan. Stem Cells and Development. 13, 22.

[157] Billy F., Clairambault J., Fercoq O. Optimisation of Cancer Drug Treatments Using Cell Population Dynamics. In: Ledzewicz U, Schättler H, Friedman A, Kashdan E, editors. Mathematical Methods and Models in Biomedicine: Springer New York; 2013. p. 265-309.

[158] Singhania R., Sramkoski R.M., Jacobberger J.W., Tyson J.J. 2011 A Hybrid Model of Mammalian Cell Cycle Regulation. PLoS Comput Biol. 7, e1001077. (10.1371/journal.pcbi.1001077)

[159] Liu Z., Pu Y., Li F., Shaffer C.A., Hoops S., Tyson J.J., Cao Y. 2012 Hybrid modeling and simulation of stochastic effects on progression through the eukaryotic cell cycle. The Journal of Chemical Physics. 136, -. (doi:http://dx.doi.org/10.1063/1.3677190)

[160] Kolewe M.E., Roberts S.C., Henson M.A. 2012 A population balance equation model of aggregation dynamics in Taxus suspension cell cultures. Biotechnology and Bioengineering. 109, 472-82. (10.1002/bit.23321)

[161] Elowitz M.B., Levine A.J., Siggia E.D., Swain P.S. 2002 Stochastic Gene Expression in a Single Cell. Science. 297, 1183-6. (10.1126/science.1070919)

[162] Stewart M.H., Bosse M., Chadwick K., Menendez P., Bendall S.C., Bhatia M. 2006 Clonal isolation of hESCs reveals heterogeneity within the pluripotent stem cell compartment. Nat Meth. 3, 807-15. (http://www.nature.com/nmeth/journal/v3/n10/suppinfo/nmeth939_S1.html)

[163] Sigal A., Milo R., Cohen A., Geva-Zatorsky N., Klein Y., Liron Y., Rosenfeld N., Danon T., Perzov N., Alon U. 2006 Variability and memory of protein levels in human cells. Nature. 444, 643-6. (http://www.nature.com/nature/journal/v444/n7119/suppinfo/nature05316_S1.html)

[164] Scott R.E., Hoerl B.J., Wille J.J., Florine D.L., Krawisz B.R., Yun K. 1982 Coupling of proadipocyte growth arrest and differentiation. II. A cell cycle model for the physiological control of cell proliferation. The Journal of Cell Biology. 94, 400-5. (10.1083/jcb.94.2.400)

[165] Deasy B.M., Jankowski R.J., Payne T.R., Cao B., Goff J.P., Greenberger J.S., Huard J. 2003 Modeling Stem Cell Population Growth: Incorporating Terms for Proliferative Heterogeneity. STEM CELLS. 21, 536-45. (10.1634/stemcells.21-5-536)

[166] Lei J., Levin S.A., Nie Q. 2014 Mathematical model of adult stem cell regeneration with cross-talk between genetic and epigenetic regulation. Proceedings of the National Academy of Sciences. (10.1073/pnas.1324267111)

[167] Gregory C.A., Singh H., Perry A.S., Prockop D.J. 2003 The Wnt Signaling Inhibitor Dickkopf-1 Is Required for Reentry into the Cell Cycle of Human Adult Stem Cells from Bone Marrow. Journal of Biological Chemistry. 278, 28067-78. (10.1074/jbc.M300373200)

[168] Li L., Clevers H. 2010 Coexistence of Quiescent and Active Adult Stem Cells in Mammals. Science. 327, 542-5. (10.1126/science.1180794)

[169] Dolezalova D., Mraz M., Barta T., Plevova K., Vinarsky V., Holubcova Z., Jaros J., Dvorak P., Pospisilova S., Hampl A. 2012 MicroRNAs Regulate p21Waf1/Cip1 Protein Expression and the DNA Damage Response in Human Embryonic Stem Cells. STEM CELLS. 30, 1362-72. (10.1002/stem.1108)

[170] Neganova I., Vilella F., Atkinson S.P., Lloret M., Passos J.F., von Zglinicki T., O'Connor J.-E., Burks D., Jones R., Armstrong L., et al. 2011 An Important Role for CDK2 in G1 to S Checkpoint Activation and DNA Damage Response in Human Embryonic Stem Cells. STEM CELLS. 29, 651-9. (10.1002/stem.620)

[171] Bárta T., Vinarský V., Holubcová Z., Doležalová D., Verner J., Pospíšilová Š., Dvořák P., Hampl A. 2010 Human Embryonic Stem Cells Are Capable of Executing G1/S Checkpoint Activation. STEM CELLS. 28, 1143-52. (10.1002/stem.451)

[172] Bertoli C., Skotheim J.M., de Bruin R.A.M. 2013 Control of cell cycle transcription during G1 and S phases. Nat Rev Mol Cell Biol. 14, 518-28. (10.1038/nrm3629)

[173] Singh A.M., Dalton S. 2009 The Cell Cycle and Myc Intersect with Mechanisms that Regulate Pluripotency and Reprogramming. Cell Stem Cell. 5, 141-9. (http://dx.doi.org/10.1016/j.stem.2009.07.003)

[174] Stead E., White J., Faast R., Conn S., Goldstone S., Rathjen J., Dhingra U., Rathjen P., Walker D., Dalton S. 2002 Pluripotent cell division cycles are driven by ectopic Cdk2, cyclin A/E and E2F activities. Oncogene. 21, 14.

[175] Faast R., White J., Cartwright P., Crocker L., Sarcevic B., Dalton S. 2004 Cdk6-cyclin D3 activity in murine ES cells is resistant to inhibition by p16INK4a. Oncogene. 23, 491-502.

[176] Filipczyk A.A., Laslett A.L., Mummery C., Pera M.F. 2007 Differentiation is coupled to changes in the cell cycle regulatory apparatus of human embryonic stem cells. Stem Cell Research. 1, 45-60. (10.1016/j.scr.2007.09.002)

[177] Conklin J.F., Baker J., Sage J. 2012 The RB family is required for the self-renewal and survival of human embryonic stem cells. Nat Commun. 3, 1244. (http://www.nature.com/ncomms/journal/v3/n12/suppinfo/ncomms2254_S1.html)

[178] Task K., Jaramillo M., Banerjee I. 2012 Population Based Model of Human Embryonic Stem Cell (hESC) Differentiation during Endoderm Induction. PLoS ONE. 7, e32975. (10.1371/journal.pone.0032975)

[179] Phillips B.W., Hentze H., Rust W.L., Chen Q., Chipperfield H., Tan E., Abraham S., Sadasivam A., Soong P.L., Wang S.T., et al. 2007 Directed Differentiation of Human Embryonic Stem Cells into the Pancreatic Endocrine Lineage. Stem Cells and Development. 16, 18.

[180] Touboul T., Hannan N.R.F., Corbineau S., Martinez A., Martinet C., Branchereau S., Mainot S., Strick-Marchand H., Pedersen R., Di Santo J., et al. 2010 Generation of functional hepatocytes from human embryonic stem cells under chemically defined conditions that recapitulate liver development. Hepatology. 51, 1754-65. (10.1002/hep.23506)

[181] Kanai-Azuma M., Kanai Y., Gad J.M., Tajima Y., Taya C., Kurohmaru M., Sanai Y., Yonekawa H., Yazaki K., Tam P.P.L., et al. 2002 Depletion of definitive gut endoderm in Sox17-null mice. Development. 129, 13.

[182] McGrath K.E., Koniski A.D., Maltby K.M., McGann J.K., Palis J. 1999 Embryonic Expression and Function of the Chemokine SDF-1 and Its Receptor, CXCR4. Developmental Biology. 213, 442-56. (DOI: 10.1006/dbio.1999.9405)

[183] Roeder I., Loeffler M. 2002 A novel dynamic model of hematopoietic stem cell organization based on the concept of within-tissue plasticity. Experimental Hematology. 30, 853-61. (Doi: 10.1016/s0301-472x(02)00832-9)

[184] Nelson T.J., Faustino R.S., Chiriac A., Crespo-Diaz R., Behfar A., Terzic A. 2008 CXCR4+/FLK-1+ Biomarkers Select a Cardiopoietic Lineage from Embryonic Stem Cells. STEM CELLS. 26, 1464-73. (10.1634/stemcells.2007-0808)

[185] Rodaway A., Patient R. 2001 Mesendoderm: An ancient germ layer? Cell. 105, 169-72.

[186] Degasperi A., Gilmore S. Sensitivity Analysis of Stochastic Models of Bistable Biochemical Reactions. In: Bernardo M, Degano P, Zavattaro G, editors. Formal Methods for Computational Systems Biology: Springer Berlin / Heidelberg; 2008. p. 1-20.

[187] Freedman D., Diaconis P. 1981 On the histogram as a density estimator:L2 theory. Probability Theory and Related Fields. 57, 453-76. (10.1007/bf01025868)

[188] McLean A.B., D'Amour K.A., Jones K.L., Krishnamoorthy M., Kulik M.J., Reynolds D.M., Sheppard A.M., Liu H., Xu Y., Baetge E.E., et al. 2007 Activin A Efficiently Specifies

Definitive Endoderm from Human Embryonic Stem Cells Only When Phosphatidylinositol 3-Kinase Signaling Is Suppressed. STEM CELLS. 25, 29-38. (10.1634/stemcells.2006-0219)

[189] Yusuf F., Rehimi R., Dai F., Brand-Saberi B. 2005 Expression of chemokine receptor CXCR4 during chick embryo development. Anatomy and Embryology. 210, 35-41. (10.1007/s00429-005-0013-9)

[190] Takenaga M F.M., Hori Y. 2007 Regulated Nodal signaling promotes differentiation of the definitive endoderm and mesoderm from ES cells. Journal of Cell Science. 120, 13.

[191] Tada S., Era T., Furusawa C., Sakurai H., Nishikawa S., Kinoshita M., Nakao K., Chiba T., Nishikawa S.-I. 2005 Characterization of mesendoderm: a diverging point of the definitive endoderm and mesoderm in embryonic stem cell differentiation culture. Development. 132, 4363-74. (10.1242/dev.02005)

[192] Loose M., Patient R. 2004 A genetic regulatory network for Xenopus mesendoderm formation. Developmental Biology. 271, 467-78. (DOI: 10.1016/j.ydbio.2004.04.014)

[193] Chng Z., Teo A., Pedersen R.A., Vallier L. 2010 SIP1 Mediates Cell-Fate Decisions between Neuroectoderm and Mesendoderm in Human Pluripotent Stem Cells. Cell Stem Cell. 6, 59-70. (DOI: 10.1016/j.stem.2009.11.015)

[194] Blais A., Dynlacht B.D. 2005 Constructing transcriptional regulatory networks. Genes and Development. 19, 13.

[195] Blum B., Benvenisty N. The Tumorigenicity of Human Embryonic Stem Cells. In: George FVW, George K, editors. Advances in Cancer Research: Academic Press; 2008. p. 133-58.